Dealing with relatives:

Population-scale pedigrees in human genetics

Dominic Nelson

Department of Human Genetics

McGill University, Montréal, Canada

May 2020

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Contents

Abstract

All of humanity, and indeed all living things, are connected by a complex web of ancestral relationships. How much of this shared pedigree is known, how much can be inferred, and how completely it can be described, are important factors in interpreting present-day genetic diversity and discovering genetic contributions to disease. The challenges of using pedigree information are twofold. First, in human populations, genealogical data is often unavailable or incomplete. Second, when such datasets are available, there are significant computational challenges to using them. In both cases the response has often been to use approximate models of pedigree structure, which removes the need for complete historical data and simplifies analysis. But genetic cohorts are growing, and increased sample sizes necessitate more detailed models of relatedness.

The work presented here represents a first step in explicitly modeling the relatedness of millions of individuals, and using it to understand the fine-scale diversity of whole populations, as well as the genetic architecture of rare diseases. First we developed a software package, ISGen (Importance Sampling in Genealogies) for inferring the ancestral origin of rare alleles in population-scale pedigrees, as well as their distribution among present-day individuals, in a fully Bayesian framework. Similar analysis had never been

possible before at that scale, and as a proof-of-concept we analyzed the distribution of the allele causing Chronic Atrial and Intestinal Dysrhythmia (CAID) within the province of Quebec. Using observations of only 11 patients and 4 heterozygous carriers of the allele, we estimated the allele frequency in 26 geographic regions of the province, finding the highest expected frequency in Charlevoix, where no carriers had yet been observed.

Despite its value for supporting the design of screening programs for rare diseases, ISGen is limited to rare alleles and a single locus. Extending to whole-genome inference is a daunting challenge, and can be aided significantly by high-quality data with a known transmission history. Simulations are a natural source of such data, yet the most efficient simulation frameworks were based on coalescent models with known biases for long genomic regions and large sample sizes. We developed two tools to address this issue. One is an extension to the state-of-the-art `msprime` coalescent simulator which allows simulations under a Wright-Fisher model. This extension makes use of the highly efficient algorithms and data structures of `msprime`, and we showed more realistic relatedness among simulated individuals as well as improved performance over coalescent simulations at whole-genome scale. We then continued to develop this extension to allow simulations to be performed within a predefined pedigree, and when used with the population-scale genealogy of Quebec, Canada we showed that simulated cohorts captured much of the structure of a real dataset from the same population, and significantly more than comparable simulations of a randomly-mating population. To show the potential of large-scale pedigree simulations more generally, we further described how pedigree simulations can aid in the design of sequencing cohorts for imputation, and be used to detect rare-variant associations while explicitly accounting for relatedness among cohort members.

Résumé

Toute l'humanité, et en fait tous les êtres vivants, sont reliés par un réseau complexe de relations ancestrales. La proportion de ce pedigree qui est connue, celle qui peut être apprise, et notre capacité à le décrire demeurent des facteurs essentiels pour interpréter la diversité génétique actuelle et découvrir les contributions génétiques à la maladie. Les défis liés à l'utilisation des informations généalogiques sont doubles. Premièrement, dans les populations humaines, les données généalogiques sont souvent incomplètes. Deuxièmement, lorsque de tels ensembles de données sont disponibles, leur utilisation pose d'importants défis informatiques. Dans les deux cas, la réponse a souvent été d'utiliser des modèles approximatifs de structure généalogique, ce qui élimine le besoin de données historiques complètes et simplifie l'analyse. Mais les cohortes génétiques se développent et l'augmentation de la taille des échantillons nécessite des modèles de parenté plus détaillés.

Les travaux présentés ici représentent une première étape dans la modélisation explicite de la parenté de

millions d'individus et leur utilisation pour comprendre la diversité à petite échelle de populations entières, ainsi que l'architecture génétique des maladies rares. Nous avons d'abord développé un logiciel, ISGen (Importance Sampling in Genealogies) pour déduire l'origine ancestrale des allèles rares dans les pedigrees à l'échelle de la population, ainsi que leur distribution parmi les individus actuels, dans un cadre entièrement bayésien. Une analyse similaire n'avait jamais été possible auparavant à cette échelle, et comme preuve de concept, nous avons analysé la distribution de l'allèle responsable de la dysrythmie intestinale et auriculaire chronique (DIAC) dans la province de Québec. En utilisant les observations de seulement 11 patients et 4 porteurs hétérozygotes de l'allèle, nous avons estimé la fréquence des allèles dans 26 régions géographiques de la province, trouvant la fréquence la plus élevée attendue à Charlevoix, où aucun porteur n'avait encore été observé.

Malgré sa valeur pour soutenir la conception de programmes de dépistage des maladies rares, ISGen est limité à des allèles rares et à un seul locus. S'étendre à l'inférence du génome entier est un défi de taille et peut être considérablement aidé par des données de haute qualité avec un historique de transmission connu. Les simulations sont une source naturelle de telles données, mais les cadres de simulation les plus efficaces étaient basés sur des modèles coalescents avec des biais connus pour de longues régions génomiques et de grandes tailles d'échantillons. Nous avons développé deux outils pour répondre à ce problème. L'une est une extension du simulateur de coalescence de pointe `msprime` qui permet de simuler sous un modèle de Wright-Fisher. Cette extension permet l'utilisation des algorithmes et les structures de données efficaces de `msprime`, et nous avons démontré une relation plus réaliste entre les individus simulés ainsi que des améliorations de performances par rapport aux simulations coalescentes à l'échelle du génome entier. Nous avons ensuite développé cette extension pour effectuer des simulations dans un pedigree prédéfini, dont celui de la population du Québec, nous avons montré que les cohortes simulées capturaient une grande partie de la structure d'un ensemble de données réel de la même population, et bien plus que des simulations comparables d'une population à accouplement aléatoire. Pour montrer le potentiel des simulations généalogiques à grande échelle de manière plus générale, nous avons décrit plus en détail comment les simulations généalogiques peuvent aider à la conception de cohortes de séquençage pour l'imputation et être utilisées pour détecter des associations de variantes rares tout en tenant explicitement compte de la parenté entre les membres de la cohorte.

List of Abbreviations

ARG - Ancestral Recombination Graph

CAID - Chronic Atrial and Intestinal Dysrhythmia

DSMC - Discretized Sequentially Markov Coalescent

GWAS - Genome-Wide Association Study

HMM - Hidden Markov Model

IBD - Identity/Identical By Descent

IRB - Institutional Review Board

ISGen - Importance Sampling in Genealogies

LMM - Linear Mixed Model

MCMC - Markov Chain Monte Carlo

MRCA - Most-Recent Common Ancestor

OMIM / MIM - Online Mendelian Inheritance in Man

PCA - Principle Component Analysis

SMC - Sequentially Markovian Coalescent

SNP - Single Nucleotide Polymorphism

TDT - Transmission Disequilibrium Test

UMAP - Uniform Manifold Approximation and Projection for Dimension Reduction

dbSNP - Single Nucleotide Polymorphism Database

List of Figures

List of Tables

Acknowledgements

This thesis was supported by generous funding from the Reseau de Médecine Génétique Appliqué, the Fonds de Recherche du Québec - Nature et Technologies, and the Queen Elizabeth II Diamond Jubilee Scholarship Program.

A huge thanks to my supervisor, Dr. Simon Gravel, for his support throughout my studies. His enthusiasm has been thoroughly contagious, and he was equally at ease whether guiding me through highly technical proofs or showing honest compassion when I faced challenges of any kind. I am truly lucky to have had such mentorship.

I would also like to thank my supervisory committee members, Celia Greenwood and David Stephens, for their guidance and encouragement over the years. Many thanks as well to all the collaborators I have had the pleasure of working with. In particular I would like to thank Jerome Kelleher for sharing his extensive skills and knowledge with me, and for his patience and good humour while he did so.

Friends and family have been absolutely essential to my completion of this work. You kept me grounded and motivated, and I hope you know that you are loved and appreciated.

Format of the Thesis

This thesis is presented in manuscript-based format. The studies described were performed under the supervision of Dr. Simon Gravel, and address two major challenges arising from modelling the relationship between population-scale pedigree structure and present-day genetic variation: performing rigorous yet computationally-feasible inference in pedigrees containing millions of individuals, and simulating large whole-genome cohorts with realistic patterns of relatedness, to better understand the origins of present-day diversity, and the genetic components of rare diseases.

Chapter 1 is a general introduction giving an overview of the motivational questions and pre-existing tools with which to approach them. Chapter 2 presents a new method for estimating regional frequencies of rare disease-causing alleles, and was published in the *American Journal of Human Genetics*. Chapter 3 describes an extension to a state-of-the-art simulation software which addresses biases in pairwise relatedness among simulated genomes, and has been published in *PLOS Genetics*. Chapter 4 is a manuscript in preparation, which describes a method for simulating within a predefined pedigree in `msprime`, and the unique applications that are possible with such a simulation tool.

Contribution to Original Knowledge

The work described in this thesis presents tools for understanding and representing the relationship between ancestral genealogical relationships and present-day genetic variation.

Chapter 2 introduces a new tool, called Importance Sampling in Genealogies (ISGen), which estimates regional frequencies of rare alleles within large genealogies. Needing only a small number of observations of the variant, ISGen can infer regional allele frequencies within genealogies of millions of individuals, a scale which was computationally prohibitive for previous methods, and which can aid in the planning and implementation of population-wide screening for rare genetic diseases.

Chapter 3 presents a new framework for simulating many whole genomes with realistic pairwise relatedness. As genetic datasets increase in size, simulations are an important tool for understanding observed genetic variation, but are difficult to generate efficiently. The `msprime` software package improved simulation speed by several orders of magnitude, but the coalescent model it implements is biased for large sample sizes and long genomic regions. We extended `msprime` with a Wright-Fisher model, which does not suffer from the same biases, and show that it produces more realistic pairwise relatedness and is more efficient than `msprime`'s coalescent simulations when simulating whole genomes.

Chapter 4 further extends `msprime`, this time to allow simulations within a predefined pedigree, which can

be as large as several million individuals. Beyond what is possible with the Wright-Fisher model, this allows simulations to capture not only realistic pairwise relatedness, but also a realistic distribution of relatedness among simulated individuals. When using real pedigrees, simulations can be used as a null model to evaluate the significance of observed allele-sharing patterns, fully incorporating complex relatedness between known carriers, aiding in the mapping of rare disease-causing genes. The method opens up many other possible applications, such as evaluating genome-wide significance thresholds in various cohort compositions, or a comparison of census and effective population sizes within real populations.

Contributions of the Authors

Chapter 2

This work was done with the following colleagues and collaborators: *Claudia Moreau, Marianne de Vriendt, Yixiao Zeng, Christoph Preuss, Hélène Vezina, Emmanuel Milot, Gregor Andelfinger, Damian Labuda, and Simon Gravel.* The recruitment of CAID patients and question of finding the ancestral origin of the mutation are due to CM, CP, HV, EM, GA, and DL. SG and I conceived of the importance-sampling method, and the method for estimating expected allele frequencies conditional on allele trajectories through the genealogy. I implemented both. SG, YZ, MV and I improved the original importance-sampling scheme to improve sampling efficiency. I wrote the manuscript with the help of SG, and CM. All authors edited and approved the manuscript.

Chapter 3

This work was done with the following colleagues and collaborators: *Jerome Kelleher, Aaron P Ragsdale, Claudia Moreau, Gil McVean, and Simon Gravel.* SG, GM, JK and I conceived the project, and JK and I implemented the Wright-Fisher extension. SG, JK, APR and I validated the simulations, and APR compared simulated and expected LD. CM curated the Genizon genotype data, performed IBD detection within the dataset. I wrote the manuscript with the help of SG and JK, and all authors edited and approved the manuscript.

Chapter 4

This work was done with the following colleagues and collaborators: *Jerome Kelleher, Luke Anderson-Trocmé, Aaron P Ragsdale, Alexandre Bureau, and Simon Gravel.* SG, AB and I conceived of the project, and JK and I implemented the pedigree-simulation framework in `msprime`. LAT and I compared simulated

genomes to real data, while APR and I validated simulations within synthetic pedigrees. JK, LAT, APR, SG and I validated the implementation. I wrote the manuscript with SG, with guidance from SG, JK, LAT, APR, and AB.

Chapter 1

Introduction and Literature Review

1.1 Genealogies, family studies, and limitations

1.1.1 Overview

The origins of genetic diversity are inseparable from the genealogical process linking all living things. Individual relatedness, family structure, genetic isolation and speciation, migration, selection, drift - all take place within a complex pedigree through which these natural processes are acted out. Even the understanding of the gene as the unit of inheritance, beginning with Mendel [1], relied on simple parent-child genealogies to match inheritance with observed phenotypes.

In this thesis we explore ways in which pedigrees are used to expand our understanding of the sources of genetic variation, its distribution through a real population, and its connection to disease. We present a collection of tools we have developed to tackle the computational challenges of working with large pedigrees, starting by tracing single allele histories, and ending with whole genomes.

1.1.2 Associations and Significance

Interpreting patterns in genetic data requires a model to interpret them. For example, alleles at higher frequency among individuals with a certain disease may be a result of random variation rather than any causal relationship. In order to separate true associations from spurious ones, we first specify the expectation under a null model of no association, and test whether observations are significantly different than the null. Pedigrees provide a powerful framework for such testing, since directly modelling ancestral relationships guards against confounding factors such as population structure.

For example, a common test for disease associations before the widespread availability of genome-wide data was the transmission-disequilibrium test, or TDT. To perform a simple TDT, a collection of parent-child trios are genotyped at a candidate locus, where children are labelled phenotypically as either affected or unaffected. Within each trio, if there is no true association between the allele and phenotype, the distribution of alleles among the children will simply follow Mendel's laws. Conversely for an allele which does have a true association to the phenotype, we expect it to be seen more frequently among affected children, which

we claim to be the case if we can statistically reject Mendelian inheritance at a given p-value threshold.

In the case of the TDT, the null model is simple to describe, and the significance of departures from it is straightforward to evaluate. This is not always the case. In an extended pedigree the null model is more complex, since their may be a higher number of meioses separating observed genotypes, and thus more possible recombination events to consider under the null hypothesis. As pedigree size increases, computational costs increase exponentially [2], quickly exceeding the capability of analytical tools designed for pedigrees of a few dozen individuals. Approximate methods, such as those based on Markov Chain Monte Carlo (MCMC), which sample possible inheritance paths through the pedigree can offer significant speed increases but, to our knowledge, have not been used on pedigrees containing more than a few thousand individuals [3].

1.1.3 From linkage studies to GWAS

As the cost of sequencing technology decreased, large cohort sizes were no longer prohibitively expensive. Besides the computational challenges of large-scale linkage studies, pedigree data is not widely available, limiting the possible applications to families with extended pedigrees, or members of one of a handful of population-scale pedigrees available around the world (see below for examples of such datasets).

A response to these challenges was the rapid growth of genome-wide association studies (GWAS) as a method for discovering new disease associations [4, 5]. With a vast amount of data available, the exact relationships between genotyped individuals can remain unknown, as the signal is strong enough to be detectable with a more approximate model. Rather than considering a specific transmission models, GWAS look for associations between phenotypes and genotypes. To account for the large number of tests being performed and control the false positive rate, is is customary to use a stringent threshold for p-values, often 5×10^{-8} [6, 7, 8]. This particular value was determined under the assumption that the cohort is composed of largely unrelated individuals.

This assumption is difficult to justify, however, as GWAS cohorts grow in size. As an illustration of the problem, Shchur et al. [9] calculated the expected number of $n-th$ degree cousins in a sample of size K drawn from a randomly-mating population of size N. When $K/N = 0.2$, we expect approximately 55% of samples to have a first cousin in the cohort, and 95% to have a second cousin. Similar results have been seen in real cohorts, with Henn et al. [10] finding that in a sample of 1000 individuals in the 23andMe European dataset, 90% were predicted to have at least one cousin in the 2nd - 9th degree, a number which climbed to 99% with a sample size of 5,000. In particular, 5000 3rd-degree and 30,000 4th-degree cousin pairs were found.

For methods which rely on filtering out related individuals, this is clearly an obstacle to scaling up cohort

sizes. The more individuals to be included, the greater their likelihood of having a relative already in the cohort, leading to wasted sequencing effort. To compare the number of relatives expected in a smaller sample, among 406 individuals, Athanasiadis et al. [11] found 3 pairs of first cousins and 1 pair of second cousins, which also fits the expectations of the model of Shchur et al. In this case filtering out relatives would not lead to a large loss of power, showing how directly modelling relatives was not necessary until large cohorts became available.

Tools do exist for association testing with awareness of the relatedness between cohort individuals. The popular association-testing software BOLT-LMM for instance [12] computes a relatedness matrix for the cohort which is used as a covariate during linear mixed-model (LMM) regression. This controls for pairwise relatedness but still misses some of the correlation structure imposed by the population pedigree, especially in the case of rare alleles shared among relatives. Since fewer ancestors carried them, rare alleles have fewer paths through the population pedigree leading to present-day individuals, so the specific path they took becomes significant.

Taking an extended family as an example, the likelihood of each family member being a carrier is not independent. Suppose a heterozygous carrier inherited an allele through their maternal side: This would imply that one of their maternal grandparents must have been a carrier, increasing the likelihood that a cousin on their mother's side is also a carrier. This fine-grained correlation structure is not captured by the relatedness matrix, limiting the ability of tools such as BOLT to be applied to rare allele sharing among family members.

1.1.4 Pedigrees in population genetics

Present-day genetic variation is a product of the population pedigree, a fixed yet generally unknown parameter [13]. The transmission-disequilibrium test described above shows that knowledge of even a small portion of the population pedigree, down to parent/child relationships, allows for sufficiently powerful statistical tests to detect disease associations. However even without knowledge of the exact pedigree, modelling assumptions about its general structure allow powerful inferences to be made about the origin of observed genetic diversity. Pedigree models and their assumptions underlie a vast array of results in population genetics. For the purposes of the present work, we will examine them in the context of genetic simulations.

One of the first systematic attempts to model pedigrees and their effects on genetic variation was done by Wright [14] and Fisher [15]. The Wright-Fisher model is foundational to much of population-genetic analysis, forming the basis of a wide range of theoretical work. In its backwards-time formulation it models genealogies as being composed of non-overlapping generations of randomly-mating individuals, where offspring are

generated by drawing gametes from parents in the previous generation. In the single-locus case, the state of the population at any point in time is therefore fully described by the allele frequency alone, and by the distribution of haplotype frequencies in the multi-locus case.

We focus here on the neutral Wright-Fisher model, which allows parents to be chosen with uniform probability from the previous generation, independent of parental genotypes. If the population contains i copies of an allele at one point in time, the probability that it contains j copies in the next generation is given by the probability that j offspring have parents who are carriers, and the remainder have non-carrier parents. Combining this with the number of possible parent-child combinations in which this outcome is possible gives, for a haploid population of N individuals,

$$P_{ij} = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{1}{N}\right)^{N-j}$$

for $0 \leq i, j \leq N$ [16]. This lack of dependence on previous population states means the Wright-Fisher model can be modelled as a Markovian processes with frequencies of 0 and 1 being absorbing states. This leads to some fundamental insights into the nature and evolution of genetic variation, such as time to fixation, heterozygosity, mutation-drift balance, and mutation-selection balance. The interested reader can refer to a wide range of literature for more details ([17, 18, 19, 16] among others).

While the Wright-Fisher simplifies many properties of real pedigrees (more on this in Chapter 3), it remains challenging to study analytically. A more tractable alternative is coalescent theory, a theoretical framework introduced by Kingman [20], Hudson [21], and Tajima [22], which makes simplifying assumptions about the Wright-Fisher model to facilitate analysis.

Here we discuss Kingman's model as generalized by Hudson to account for recombination. First, this coalescent model assumes that sample sizes are small relative to effective population sizes. This means we can assume that sampled individuals are not closely related, so patterns of diversity will be relatively unaffected by demographic events in the recent past. Second, the coalescent allows only a single ancestral event at any given time point, which can be either a recombination event or a coalescence event. Single recombination events mean that back-and-forth recombination is not modelled, and similarly only pairwise coalescence events are described. For small sample sizes we would expect very few multiple-merger events, so this is a good approximation, and assuming the genomic region of interest is sufficiently short, back-and-forth recombination can be neglected. This means each recombination event creates a new, independent lineage, which traces a new path through the population pedigree.

However, there are conceptual limitations to this representation of the population pedigree. The population pedigree is not in fact random [23, 24], since it is simply a representation of the parent-offspring

relationships that gave rise to all contemporary individuals, which, being events in the past, are fixed.

When the coalescent draws genealogies, they are not conditioned on past coalescence or common-ancestor events, effectively creating a new population pedigree at each locus. The problem becomes most clear when looking at the recent past. If two sampled individuals share a recent common ancestor, this should properly be considered when drawing genealogies at all loci. If for simplicity we consider siblings, we would expect many simultaneous common-ancestry events, across all chromosomes, to occur within their parents. Since the coalescent does not describe such simultaneous events, even siblings will only share a single common-ancestry event, regardless of how long a genomic region is considered. As we will see in Chapter 2, this can lead to substantial biases in relatedness among simulated present-day individuals.

1.2 Large genealogical datasets

Genealogies now reach into the millions of individuals. This is a tremendous resource for medical genetics, but practical use is limited by computational challenges. These challenges are imposing, but the benefits motivate us to confront them, and we expect the availability of these large datasets to grow as time goes on. After reviewing some of the largest genealogical datasets currently available, we turn our attention to a relatively simple scenario, as a first step in developing computationally feasible tools for inference within pedigrees of millions of individuals.

1.2.1 Large datasets

One large genealogy is the deCODE database, which has been used extensively in medical and population genetics `https://www.decode.com/publications/`. Based in Iceland, nearly all of the present-day population of 330,000 is contained within the genealogy, and nearly all of their ancestors back to the year 1650 [25]. Approximately half of present-day individuals (numbering roughly 150,000) have been genotyped, and around 40,000 have had whole-genome sequencing [25]. Despite a relatively small population size, inbreeding overall is low, and there is a low rate of autosomal recessive disorders [26]. The deCODE database, with its extensive genetic and genealogical data, supports studies on selection, diabetes, heart disease, cancer, and Alzheimer's disease [27, 28, 29].

Another large genealogy is the Utah Population Database, composed of 11 million individuals mostly from the 18th century to the present, but with some genealogies reaching 17 generations into the past [30, 31]. It has been constructed by combining records from multiple sources: birth, death, and marriage certificates, original family history records, US census data, as well as more recent medical and insurance records. The broad range of data available for many individuals, especially contemporary ones, supports a variety of active

research areas, including mental health, Alzheimers, cancer, and longevity [32, 33, 34, 35, 36].

The BALSAC genealogical database [37], used extensively in the work presented here, is constructed from over 3 million historical records, mostly birth, death, and marriage certificates, altogether encompassing over 5 million individuals from the French-Canadian population of Quebec. The genealogy extends a maximum depth of 17 generations, and most present-day individuals can trace at least one lineage 12 generations in the past. The accuracy of links within the BALSAC database was examined in [38, 39], and errors are quite low, with approximately 1% false paternity detected. The depth and completeness of the database make it particularly well-suited for studies on selection, admixture, spatial demography, and population-scale risk analysis, as well as rare cancers and cardiac and neurological disorders [40, 41, 42, 43, 44, 45, 46]. Further applications will be possible as more genetic and phenotype data is integrated from the Genizon and CARTaGENE biobanks [47, 48].

Large pedigrees have also been crowd-sourced, such as the fully-connected pedigree of 13 million individuals constructed from genealogical data input by users of the website `Geni.com` [49]. This was built from 43 million profiles in which users had added genealogical connections to other users in the database. To evaluate the accuracy of the resulting pedigrees, invalid topologies (such as an individual having more than two parents) were pruned with an automated pipeline, and tree structures were evaluated against both mitochondrial DNA (211 lineages) and Y-chromosome short tandem repeat haplotypes (27 lineages). These resulted in an estimated nonmaternity rate of 0.3% per meiosis, and a nonpaternity rate of 1.9% per meiosis, matching respectively the historical adoption rate of non-relatives in the US [50], and the results of previous Y-chromosome studies [51, 52]. The genealogy has been used to study the genetic architecture of longevity, and to analyse changing familial dispersion over several centuries [49].

Large pedigrees have also been constructed by private companies. For example, Ancestry.com have a database of over 20 million genealogical records [53], although they are spread over a large number of smaller fully-connected pedigrees. These have been used to study IBD and historical migration in the United States, where the pedigrees help distinguish genetically similar groups such as Acadians and French Canadians [53]. Private datasets are generally not readily accessible to researchers, although they do suggest further demand for scalable pedigree-analysis tools.

1.2.2 ISGen

Our motivation to develop inference tools suitable for such large datasets began with the following question: given a few individuals in the genealogy known to carry a rare mutation, what can we learn about the origins and distribution of this mutation in the general population? The first manuscript presented here in Chapter

1 describes the implementation of a fully Bayesian approach to this problem, by inferring transmission histories in pedigrees of millions of individuals. This is orders-of-magnitude larger than any genealogy previously studied using a fully Bayesian framework [3].

The tool we developed for this purpose, called ISGen (Importance Sampling in Genealogies), takes as input a panel of carriers of a rare allele, and samples millions of possible transmission histories of the observed mutation. These paths are then integrated to compute both the posterior probability of each pedigree founder having introduced the mutation, as well as the expected allele frequency within arbitrary groups of pedigree probands. Regional allele frequencies, for instance, can be obtained by grouping probands by geographical location, when such data are available. These estimates can then guide the design of screening programs for rare genetic disorders.

However, ISGen can only be applied when certain assumptions hold, some of which are fairly restrictive. First, the method assumes that the allele of interest is rare enough that we can assume it was introduced into the population by a single founder. This is not unreasonable for many rare diseases, but it limits the kind of inference which can be performed. Second, ISGen only infers the trajectory of a single locus through the pedigree, again limiting possible applications and losing power gains that would come from integrating likelihoods over multiple loci. While ISGen does incorporate shared haplotype length among all observed carriers, it does not use pairwise shared haplotype length to estimate the likelihood of coalscence times between pairs of allele lineages, nor the length or distribution of shared haplotypes around other loci.

Both of these assumptions were made for the sake of computational efficiency. Assuming that all observed alleles share a common ancestor within the pedigree allows many possible inheritance trajectories to be ignored, if they violate this assumption. Compared with pure Monte Carlo simulations, convergence of the importance-sampling method is sped up by a factor of approximately 10^{30}. This efficiency means that inference within the entire population-scale pedigree is computationally feasible, but the assumptions made for those efficiency gains also limit the application of ISGen beyond the study of recent rare mutations.

In the following section we describe other tools being designed for large-scale whole-genome genealogical inference. While scaling up presents challenges to maintaining efficiency, the gain in signal from the increased information makes the use of more heuristic methods attractive. However it is particularly important to validate such heuristic methods, a process which itself presents significant computational challenges. We next describe several large-scale genealogical inference methods before moving on to simulations and validation.

1.3 Genome-wide genealogical inference

1.3.1 Overview

While ISGen proves that inference can be efficiently performed in pedigrees of millions of individuals, the assumptions it requires somewhat limit both its power and its practical use. The use of haplotype length is a straightforward way to improve the power of single-locus inference, but much greater information could be gained by incorporating the full IBD structure of the sequenced cohort. It is computationally challenging to jointly infer genealogies across long genomic regions, but there are several tools now available which confront the challenge in different ways. We give an overview of their respective strategies now.

1.3.2 ARGweaver

ARGweaver [54] uses a discretized extension of the piecewise sequentially Markov coalescent, called the discretized sequentially Markov coalescent, or DSMC, as the basis of a MCMC sampler of the ancestral recombination graph, or ARG [55]. The ARG is simply a representation of the coalescent history of all sampled loci, along with a record of all past recombination events. In the DSMC, time is discretized into K blocks, which are distributed logarithmically to allow greater granularity of structure in the recent past, where density of coalescence and recombination events is highest.

ARGweaver begins by constructing an approximate ancestral recombination graph (ARG) for all input sequences. This is constructed heuristically, starting with a single sequence, where the ARG will have only a single branch at each locus. Sequences are added one at a time, with branches fit into the constructed ARG of the previous sequences. This is not equivalent to sampling an ARG from the posterior distribution, but efficiently constructs an initial point from which to begin MCMC sampling.

Samples from the posterior ARG distribution are drawn in one of two ways. The most straightforward is to take an initial ARG, remove the external branches connecting a single sequence, and then draw reconnections for this sequence for each local tree. This continues by drawing another sequence to remove and reconnect, either randomly or by iterating through all sequences in order. This explores the desired posterior distribution, but since it redraws only external branches (those connecting to a leaf node), it is not efficient at sampling different deep structures of the ARG, since internal branches will only be redrawn after multiple external branches have been redrawn in a way which 'exposes' internal branches to become external.

This limitation is addressed by a more sophisticated sampling strategy where whole subtrees, connecting multiple sequences, are removed and resampled. Because these subtrees include internal nodes, the deep structure of the ARG is sampled much more efficiently, leading to better mixing, especially with a large

number of sequences. The difficulty with this strategy is that in general subtrees will not span the entire ARG, and internal branches may be highly constrained in their attachment points if they are to be consistent with surrounding local trees, leading to frequent resampling of the original ARG. ARGweaver is able to select subtrees for resampling which are not overly constrained in this way, and although some efficiency is sacrificed, mixing is still improved over single-sequence resampling.

1.3.3 Relate

Relate [56] takes a different approach, which begins by constructing a distance matrix for all haplotypes. This is done using an HMM similar to that used by Li and Stephens [57], but using extra information from the ancestral or derived states of each SNP to improve accuracy of estimates and speed up inference. This matrix is used to construct a gene genealogy using a hierarchical clustering method, which assumes that distances between haplotypes correspond to the order of coalescence events between them. Initial clusters are created, after which the distance matrix is updated to reflect distances between these clusters, with this process continuing until the full genealogy is constructed.

Once the tree topology is determined, mutations are mapped to tree branches. Often this mapping is unique and unambiguous, but if repeat mutations are necessary to match observed data, the mutation is mapped to the smallest number of branches necessary. Ancestral and derived states are also swapped if this results in a simplified mapping. For computational efficiency, trees are constructed by scanning along the genome until a mutation cannot be mapped to a unique branch.

Tree branch lengths are estimated using a Metropolis-Hastings MCMC algorithm with a coalescent prior and a single panmictic population, with population size estimated jointly. Coalescent rates are first estimated from branch lengths, which are then used to estimate the effective population size. This new population size is then used to update estimated branch-lengths, and the process continues until branch lengths and population size converge.

Scaling is linear in sequence length and quadratic in sample size, and can infer genealogies for 10,000 human genomes. In simulations under the coalescent with recombination, using `msprime` [58], Relate was over 4 orders of magnitude faster than ARGweaver [56].

1.3.4 `tsinfer`

Another tool, `tsinfer` [59], begins with a heuristic method for inferring ancestral haplotypes surrounding each observed variant. Each variant is assumed to be the result of a single mutation, and so can be mapped onto a unique ancestral haplotype. Taking the frequency of a focal variant as a proxy for its age, and therefore

also the age of the first ancestral carrier, the ancestral haplotype is then extended using information from sampled individuals who carry the variant. Extending in both directions, the age of subsequent variants are similarly estimated using their frequencies, and the ancestral haplotype is updated according to whether these variants are estimated to be older or younger than the original ancestor. If the variant is younger, then it cannot have been carried by the ancestor, so the ancestral haplotype is assigned a '0' at this variant locus. If the variant is older, the ancestral haplotype is assigned the most-common value among all present-day carriers of the focal variant. As the ancestral haplotype is extended, present-day individuals who no longer match the ancestral haplotype are assumed to have undergone a recombination event, and are no longer used to build the ancestral haplotype. Once more than half of carriers of the focal variant have been removed in this way, this process terminates and the inferred ancestral haplotype is considered complete.

Present-day genomes are then represented as a mosaic of these inferred ancestral haplotypes using a Li and Stephens model [57] with an added state for non-ancestral material which was not inherited by any sampled individuals, and so cannot be copied. Ancestral haplotypes are themselves modelled as mosaics of older ancestral haplotypes, a process which continues until there is a complete genealogy for each locus.

`tsinfer` is able to infer genealogies for hundreds of thousands of whole chromosomes in a few hours, albeit with somewhat large memory requirements. For example, running on chromosome 20 in the approximately 487,000 participants in UK Biobank took 3h of runtime across 40 cores, and 160GB of memory.

1.3.5 Summary

These methods show considerable promise for storing and interpreting biobank-scale data, and a useful metric for evaluating their accuracy and potential biases is to apply them to data with a known genealogy. Testing on simulated data is a natural strategy, since real genetic data with known gene genealogies is not available at a comparable scale (although interesting work has been carried out genotyping a near-complete pedigree of Florida Scrub-Jays, totalling 3,404 individuals, over a period of several decades [60]). Generating simulated data proves also to be a challenge, since state-of-the art simulation software either could not scale to the size of modern cohorts, or suffered from biases when simulating many individuals, and in particular many whole genomes.

1.4 Simulator background, scaling challenges, Wright-Fisher solution

1.4.1 `ms` and SMC approximation

Hudson's `ms` software [61], for simulating genetic variation under the coalescent with recombination, represented a significant advance in simulator efficiency. By making some simplifying assumptions - that sample size is small relative to effective population size, and that sequence length is short enough that back-and-forth recombination can be neglected - `ms` is able to simulate the genetic history of the samples back to their most-recent common ancestor at all simulated loci. These assumptions held for the scale of genetic studies at the time, and `ms` has been used in a wide range of applications, having been cited over two thousand times since publication [62, 63, 64].

But as sequencing technology became more affordable, `ms` struggled to match the scale of modern sequencing cohorts, and new simulation methods were developed. The most efficient of these were based on the sequentially Markov coalescence (SMC) model [65], where genealogies are constructed sequentially along the genome, each depending only on the genealogy at the previous locus. While these discard long-range correlations along the genome, the increased computational efficiency allows simulation of sample sizes in the tens of thousands [66].

1.4.2 `msprime`

SMC simulators remained state-of-the-art until the release of `msprime`, a new implementation of Hudson's algorithm used in `ms`, but with a new highly-efficient data structure: the *succinct tree sequence*. Before tree sequences (as we will generally refer to them here), the output of coalescent simulators such as `ms` was stored in Newick tree format, a text-based format where each marginal tree is stored in its entirety. Since most genealogical links are unchanged between nearby marginal trees, Newick tree format contains much redundancy and is accordingly slow to parse and inefficient to store [58]. A tree sequence is a binary format which removes the redundancy of Newick trees, and improves parsing and storage efficiency by several orders of magnitude. This is done by storing each node, and each edge connecting nodes, only once across all marginal trees, so that redundancy between adjacent trees adds no extra storage overhead. With simulation output containing 100,000 individuals, tree sequences in `msprime` required 40,000 times less storage space and were over a million times faster when iterating over all marginal trees [58].

1.4.3 Limitations of the coalescent

However this new efficiency exposed the limitations of the underlying coalescent model, which was developed to simulate data at a smaller scale, and exhibits biases when sample sizes are large or simulated regions are long. Genealogical distortions due to large sample sizes have been shown to affect the frequency of rare alleles [67]. Large sample sizes increase the probability of triple-mergers (or higher), which are not modelled by Hudson's coalescent. These are approximated by sequential double-mergers, but the excess of double-mergers in turn leads to an excess of doubletons in the samples. Similarly, the number of singletons is decreased relative to the Wright-Fisher model, since the leaves of the genealogy become shorter as a result [67].

While some differences were small, or at least hard to detect [24], we observed large biases in patterns of relatedness within large simulated cohorts. In order to address these biases, our strategy was to use `msprime` as a starting point, allowing us to take advantage of its high efficiency, and extend it to allow simulations at a larger scale than the coalescent model allows. Chapter 2 describes how we address these issues by returning to the model upon which coalescent theory is based: the Wright-Fisher model. The assumptions made by the coalescent make it much more accessible to mathematical analysis, and lead to a large advantage in efficiency when simulating small numbers of short genomic segments. But we describe in Chapter 2 how these advantages disappear as more and longer segments are simulated, ultimately leading to an efficiency advantage for the Wright-Fisher at biobank scales.

1.5 Wright-Fisher limitations, relatives, GWAS-scale family studies

But this is only part of the solution. Wright-Fisher simulations are an improvement, but still unrealistic for many populations. Spatial structure, inbreeding, and assortative mating are only a few examples of pedigree features that the Wright-Fisher model has limited ability to replicate. The tools for genealogical inference, described above, are meant for use on data from real populations, where these genealogical features have shaped their genetic history. Ensuring accurate inference therefore requires a more complex generative model for validation.

In GWAS, relatedness among sampled individuals must also be accounted for as cohort size continues to increase. While it has generally been accepted practice to filter out relatives from such cohorts, this is practical only when the number of such relatives is low. In large cohorts relatives are generally unavoidable, with most individuals expected to have at least a detectable cousin in the cohort. Filtering these out may drastically reduce the size of the cohort, leading to loss of power as well as wasted sequencing effort.

In Chapter 3 we present a strategy for addressing these challenges, by building a tool for efficient whole-genome simulations within arbitrary fixed pedigrees, something which has not previously been possible at large scale due to computational challenges. Having already developed a highly efficient Wright-Fisher simulator, described in Chapter 2, we build on our previous work by adapting it to arbitrary pedigrees and non-discrete generations, without sacrificing the efficiency required to generate whole-genome datasets containing hundreds of thousands of individuals and above.

With such a tool, the validation of genealogical inference is straightforward: simulations can be performed within pedigrees generated to possess specific characteristics to test the sensitivity of the inference algorithms to various pedigree structures, and real pedigrees can be used to estimate inference accuracy in real populations. While our simulation framework cannot yet capture all sources of variation, such as structural rearrangements and transposable elements, we show that simulations within a real pedigree correspond well with real data, exhibiting complex population structure which is absent from conventional Wright-Fisher simulations.

In GWAS, pedigree simulations offer a new avenue for the discovery of disease associations. When a large pedigree is available, whole-genome simulations offer a streamlined framework for testing the significance of observed patterns of allele sharing, by generating a genome-wide null model that avoids the complexity of multiple-testing corrections for single-locus statistics. Due to the real pedigree, relatedness between individuals is fully accounted for, and in particular when very large pedigrees are available this includes background levels of IBD sharing within the whole population, as well as de novo mutations. We give an example using the population-scale pedigree of the French-Canadian population of Quebec, Canada, showing how such a method can be of great use when applied to rare diseases in particular. Other applications are possible, and are discussed in Chapter 3.

1.6 Hypothesis and Objectives

The growing availability of large-scale genealogical data, particularly the BALSAC database in the province of Quebec [37], highlights a need for tools which can incorporate this data in a computationally efficient manner. One potential application is to trace genetic inheritance within a population-scale pedigree, which we hypothesize can provide improved estimation of regional carrier rates relative to kinship-based estimation. Our first objective is therefore to develop a computationally-tractable statistical method for estimating regional frequencies of rare alleles using large pedigrees.

The size of modern genetic cohorts has exposed the computational limitations of many genetics simulators. The state-of-the-art `msprime` genetics simulator has vastly improved computational efficiency but fails to

capture long-range correlations along the genome. We hypothesize that generalizing `msprime` to Wright-Fisher model will better represent human genetic diversity. Our second objective is to extend the msprime simulation package to allow simulating under a Wright-Fisher model and explore differences relative to Hudson's coalescent.

State-of-the art simulation models allow simulations under a range of broad demographic parameters, but still assume random mating within populations. We hypothesize that simulations performed within real pedigrees will allow simulated data to capture much of the variation present in real datasets. Our third objective is therefore to extend the msprime simulation package to allow simulations within a predefined pedigree, and compare simulated data to real genotype data.

Preface to Chapter 2

In this section we build a model to infer the transmission history of a single locus through a pedigree of millions of individuals, and to demonstrate that inference within pedigrees of that size is in fact computationally tractable. The work is motivated by the study of Chronic Atrial and Intestinal Dysrhythmia (CAID), a rare monogenic disease, in the French-Canadian population of Quebec, Canada. We first investigate the possibility of inferring the ancestral origin of the causal allele, by identifying the most likely founder to have introduced the allele into the population. We further use the results of this inference to estimate the regional prevalence of the disease, within any desired geographical regions, with applications to introducing or improving genetic screening for rare diseases on a per-region basis.

Chapter 2

Inferring Transmission Histories of Rare Alleles in Population-Scale Genealogies

Dominic Nelson[1], Claudia Moreau[2], Marianne de Vriendt[1, 3], Yixiao Zeng[1, 4], Christoph Preuss[2, 5], Hélène Vézina[6], Emmanuel Milot[7], Gregor Andelfinger[2], Damian Labuda[2] and Simon Gravel [1]

Published in the *American Journal of Human Genetics* 103(6): 893-906, on December 6th, 2018.

1. McGill University and Genome Quebec Innovation Centre, Montréal, QC H3A 0G1, Canada
2. Centre Hospitalier Universitaire Sainte-Justine Research Centre, Pediatrics Department, Université de Montréal, Montréal, QC, H3T 1C5, Canada
3. Biology Department, École polytechnique, 91120 Palaiseau Cedex, France
4. Lady Davis Research Institute, Jewish General Hospital, Montréal, QC, H3T 1E2, Canada
5. The Jackson Laboratory, Bar Harbor, ME 04609, USA
6. BALSAC Project, Université du Québec à Chicoutimi, Chicoutimi, QC, G7H 2B1, Canada
7. Chemistry, Biochemistry and Physics Department, and Forensic Research Group, Université du Québec à Trois-Rivières, Trois-Rivières, QC, G9A 5H7, Canada

Abstract

Learning the transmission history of alleles through a family or population plays an important role in evolutionary, demographic, and medical genetic studies. Most classical models of population genetics have attempted to do so under the assumption that the genealogy of a population is unavailable and that its idiosyncrasies can be described by a small number of parameters describing population size and mate choice dynamics. Large genetic samples have increased sensitivity to such modeling assumptions, and large-scale genealogical datasets become a useful tool to investigate realistic genealogies.

However, analyses in such large datasets are often intractable using conventional methods. We present an efficient method to infer transmission paths of rare alleles through population-scale genealogies. Based on backward-time Monte Carlo simulations of genetic inheritance, we use an importance sampling scheme to dramatically speed up convergence. The approach can take advantage of available genotypes of subsets of individuals in the genealogy including haplotype structure, as well as information about the mode of inheritance and general prevalence of a mutation or disease in the population. Using a high quality genealogical dataset of over three million married individuals in the Quebec founder population, we apply the method to reconstruct the transmission history of Chronic Atrial and Intestinal Dysrhythmia (CAID), a rare recessive disease. We identify the most likely early carriers of the mutation, and geographically map the expected carrier rate in the present-day French-Canadian population of Quebec.

Introduction

A large number of Mendelian disorders derive from well-characterized rare genetic variants [1]. Characterizing the population frequency and geographic distribution of such variants plays a central role in apportioning financial resources towards individual diagnostics, population screening and genetic counseling services [2, 3]. However, assessing regional population frequencies requires thorough clinical or genetic testing which can be costly, especially when disease mutations are rare.

Genealogical data, where available, can provide information about disease risk in untyped individuals: immediate family history is a key factor in deciding screening regimes for a range of diseases [4] such as breast cancer [5, 6, 7] and colorectal cancer [8]. Broader relatedness patterns are used to determine screening regimes for population-specific traits, especially in founder populations [4, 9, 10].

Extended family history bridges the gap between immediate family history and population-scale risk, but it is often unavailable and incomplete. Even when available, it demands careful statistical analysis. Here we are interested in using large-scale genealogies to investigate individual risk factors at the population scale,

by inferring the transmission path of disease alleles within a genealogy.

We will focus on genealogical records provided by the BALSAC [11] database, which contains 2.9 million vital event records, such as those relating to birth, death, and marriage, and consider a single connected genealogy of over 3.4 million individuals stretching from the arrival of European settlers in the Canadian province of Quebec in the 17th century, up until the present day, and spanning multiple regional founder effects [12].

Performing statistical analyses in such large genealogies is challenging. Both forward- and backward-simulations can be performed efficiently in very large genealogies [13, 14]. However, neither can be easily conditioned on observed data: forward simulations (allele dropping) are unlikely to produce the observed distribution of carriers, while unbiased backward simulations (allele climbing) are unlikely to produce plausible coalescence histories for rare variants, as we show in the Materials and Methods section below.

While many robust statistical tools exist for performing inference within genealogies, primarily for the purpose of performing linkage analysis [15, 16, 17, 18, 19, 20, 21], few are able to handle thousands of samples, let alone millions. Geyer and Thompson used a simulated tempering MCMC scheme to impute ancestral carrier status in a Hutterite genealogy with 2024 members [22]. Generalizing MCMC approaches to much larger genealogies presents formidable challenges for memory usage and convergence (E. Thompson, personal communication).

Previous work estimating prevalence using population-scale genealogies used heuristics to estimate regional prevalences across regions. For example, Chong et. al. [14] used forward simulations to estimate the distribution of allele frequencies of mutations derived from a single founder, but without taking into account specific carrier status of present individuals. Similarly, Vézina et al. [6] estimated regional prevalences of a mutation in BRCA1 in Quebec using an earlier version of the BALSAC database. They first identified a likely founder carrier of the mutation, using a heuristic based on differential genetic contribution to cases and controls, and then mapped the genetic contribution of this ancestor to each of 23 geographic regions in Quebec. Another feasible heuristic, for rare variants, is to estimate the mean kinship of individuals in a given region to known cases. Neither heuristic models correlations in genotypes among cases, which can bias estimates.

The work presented here aims to provide a more accurate and rigorous statistical framework for generating regional estimates, and more generally performing inference in very large genealogies that are being generated on academic, private, and participatory platforms [11, 23, 24, 25]. We present a general and scalable method and software package, ISGen, which uses importance sampling and careful software implementation to perform carrier risk analysis in such databases. ISGen takes as input available genotypes of specific individuals within the genealogy, including known cases, carriers, and genotyped relatives. It can use

information about population-level estimates of the carrier rate in the general population as well as haplotype sharing information. ISGen uses importance-weighted allele climbing to efficiently explore transmission history space for neutral or recessive lethal alleles. Simulations show that it can be used to estimate regional prevalences more accurately than approaches based on kinship alone.

Because ISGen computes the likelihoods of a large number of possible inheritance paths consistent with an observed set of known patients and carriers, it can also be used to compute the posterior probability that a given ancestor introduced the mutation in the population through mutation or immigration. We use this method to infer the most likely ancestral origin of a rare allele causing Chronic Atrial and Intestinal Dysrhythmia (CAID, [MIM: 616201]), a recessive disorder within the present-day population of Quebec, Canada, from among the first Europeans to settle in the area in the early 17th century. We then map the expected frequency of the allele in 23 regions of Quebec. The Materials and Methods section presents the technical details of the algorithm and implementation, as well as validation results, while the Applications section presents the analysis of the CAID allele.

Materials and Methods

Data and Initialization

ISGen explores, through Monte Carlo simulation, the set of possible genotype assignments within a genealogy that are consistent with observed genotypes and with other assumptions about the inheritance mode and ancestral frequency. At the beginning of a simulation, most genotypes are unknown (i.e., unassigned), and only the genotypes of known cases, carriers, and their relatives are set to their observed values. The genealogical relationships themselves are recorded as a table of parent-offspring triplets, as shown in Table 7.1.

Monte Carlo Simulations

After initialization, the process of allele climbing begins. We simulate the inheritance of each minor allele through either the maternal or paternal side, setting unobserved parental alleles to match those of the climbing allele. This simulated inheritance continues upwards through grandparents and more distant ancestors until reaching the 'founders' of the genealogy, i.e., individuals with one or two missing parents in the genealogy (Fig. 2.1A). In practice, because the BALSAC dataset relies on marriage records, there are no 'half-founders' with a single known parent in the genealogy, and in the following we use 'founders' to refer to individuals with no parents in the genealogy. When multiple minor allele copies are inherited from the

same individual, we say that they coalesce if they are inherited from (i.e., climb to) the same allele copy, otherwise the individual is inferred to be a homozygote.

Major and minor alleles can be treated in a symmetric manner during allele climbing. However, because the number of major allele copies in the population is usually much greater than that of minor alleles, we find it more numerically efficient to first perform allele climbing on minor alleles as outlined in this section, and then use a different procedure for estimating likelihood based on major allele carriers, which is outlined later in this section. Similarly, haplotype information is included at a later stage, and is also outlined below.

By tracing lineages of each minor allele copy through the genealogy, we define a possible allele transmission history consistent with the observed carriers. This history defines an *inheritance path*, the set of individuals either known or inferred to carry a minor allele. It is possible (indeed overwhelmingly likely) for a randomly sampled inheritance path not to have fully coalesced within the genealogy.

We focus on alleles that are rare among the founders. Specifically, we assume that the allele frequency in the ancestral population from which the founders originate is $\omega \ll \frac{1}{N_{founders}}$, where $N_{founders}$ is the number of founders, implying that the allele most likely came from a single founder. The assumption of a single origin is not central to the approach, but it simplifies the description and speeds up the inference. It is a reasonable assumption for rare diseases in small founder populations [14], but a relaxation of this assumption is outlined in the Discussion.

To compute the likelihood that ancestor a contributed the set of haplotypes c that were observed to carry the minor allele, we simply compute the proportion of simulations that coalesce from c into ancestor a. Let S be the observed event that all haplotypes in c carry the minor allele. Let Γ denote a simulated inheritance path ascending from c, and let A be a random variable representing the founder who carried the minor allele. If $\mathbb{1}_a(\Gamma)$ is the indicator function for whether Γ coalesces to founder a, and M the number of Monte Carlo iterations, we estimate the likelihood as

$$P(S|A=a) = P(\Gamma \text{ coalesces to } a) = E\big[\mathbb{1}_a(\Gamma)\big] \simeq \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}_a(\Gamma_j). \tag{2.1}$$

where the last step is a Monte Carlo integration, and Γ_j is the inheritance path constructed in simulation j, drawn from distribution $p(\Gamma_j)$ defined by the allele climbing process.

Assuming a flat prior for all ancestors a in the set $\mathcal{A}$ of all founding ancestors, Bayes theorem provides the normalized posterior probability that a is the founding carrier:

$$P(A=a|S) = \frac{P(S|A=a)P(A=a)}{\sum_{a'\in\mathcal{A}} P(S|A=a')P(A=a')} = \frac{P(S|A=a)}{\sum_{a'\in\mathcal{A}} P(S|A=a')}. \tag{2.2}$$

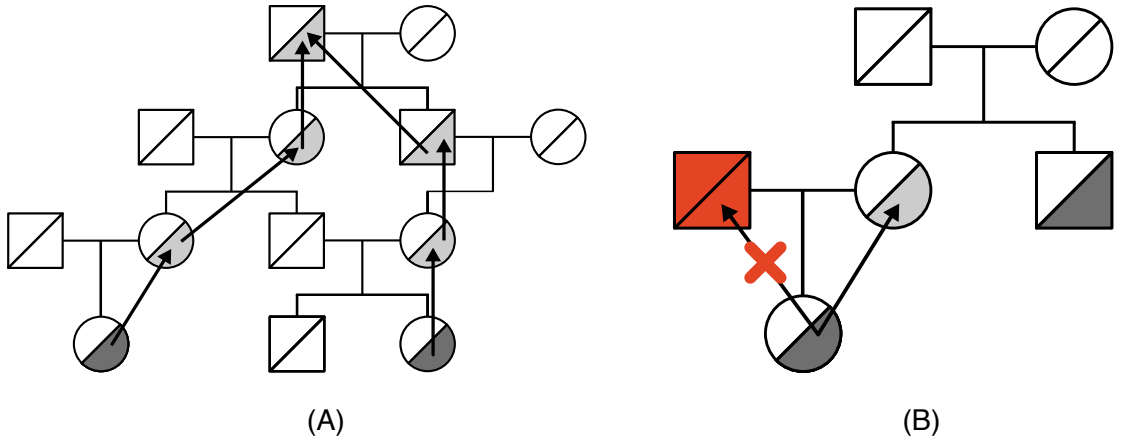


Figure 2.1: (A) Alleles are assigned to probands, and then climb up the genealogy by choosing to follow either maternal or paternal inheritance. (B) In the simplest importance sampling scheme, ISGen ensures that the red individual is never assigned an allele, since then full coalescence within the genealogy would be impossible. It adjusts the likelihood by a factor of $1/2$ to avoid biasing maximum likelihood estimate.

In practice, we perform a single Monte Carlo simulation to estimate simultaneously $P(S|A = a)$ for all ancestors a. Even then, because coalescence to a single ancestor is a very rare occurrence in a large genealogy, the majority of simulations yield $\mathbb{1}_a(\Gamma_j) = 0$ for all a and do not inform our likelihood estimate.

Importance Sampling

The Monte Carlo distribution $p(\Gamma)$ generates mostly inheritance paths with zero likelihood. To improve convergence, importance sampling uses a heuristic proposal distribution $q(\Gamma)$ to favor higher-likelihood paths. As long as we account for the over-representation of these paths, the resulting estimates are unbiased.

A simple importance sampling scheme

In the course of a simulation, it is simple to assess whether individuals in an incomplete inheritance path share a common ancestor. When simulating an allele inheritance, a simple importance sampling scheme would be to verify whether each of the maternal and paternal paths is consistent with eventual coalescence, and forbid inconsistent choices (Fig. 1B). Being 'consistent with coalescence' means sharing a common ancestor with the other lineages in the sample and, in the case of a homozygote, sharing such a common ancestor through both paternal and maternal lineages.

This defines a simple proposal distribution $q(\Gamma)$ under which all paths coalesce to a single ancestor a and

contribute to the likelihood. To obtain unbiased likelihood estimates, we need to identify the likelihood ratio $\frac{p(\Gamma)}{q(\Gamma)}$ for each sample path Γ. The Monte Carlo sampling probability for Γ is

$$p(\Gamma) = 2^{-\alpha}$$

where $\alpha = \alpha(\Gamma)$ is the number of allele transmissions in Γ. If Γ coalesces to a single ancestor a, it has a higher probability under q:

$$q(\Gamma) = 2^{-(\alpha-\beta-\gamma)}$$

where β is the number of transmissions with only one valid maternal/paternal path consistent with coalescence, and γ is the number of times a homozygote inconsistent with coalescence could have been created during the climbing process (homozygotes need a path to coalescence through both parents). Thus the likelihood ratio is

$$\frac{p(\Gamma)}{q(\Gamma)} = 2^{-\beta-\gamma}. \tag{2.3}$$

For patient panels of tens of individuals in the BALSAC genealogy, a representative histogram of values for this ratio are shown in Fig. 2.2. The importance sampling estimate of $P(S|A = a)$ is then

$$\begin{aligned} P(S|A = a) &= \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}_a(\Gamma_j)\frac{p(\Gamma_j)}{q(\Gamma_j)} \\ &= \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}_a(\Gamma_j)\ 2^{-\beta_j-\gamma_j} \end{aligned} \tag{2.4}$$

where Γ_j denotes the inheritance path drawn from q in simulation j.

This framework is flexible enough to include rather general conditions on the inheritance paths. For example, if we climb an allele known to cause a lethal recessive disease, we can ensure there are no homozygous individuals in our simulated lineages by using importance sampling to avoid simulating homozygotes altogether: we do this when applying ISGen to a lethal recessive disease in the Applications section.

We present a more elaborate importance sampling scheme below, but for clarity of exposition we use the simple scheme presented above to introduce model extensions.

Incorporating major alleles and the observed allele frequency

Through allele climbing, Equation (2.4) computes the probability that a given ancestor gave rise to specific minor alleles. However, a complete model must also take into account the distribution of major alleles. We use two approaches to model this distribution, depending on the type of information that is available.

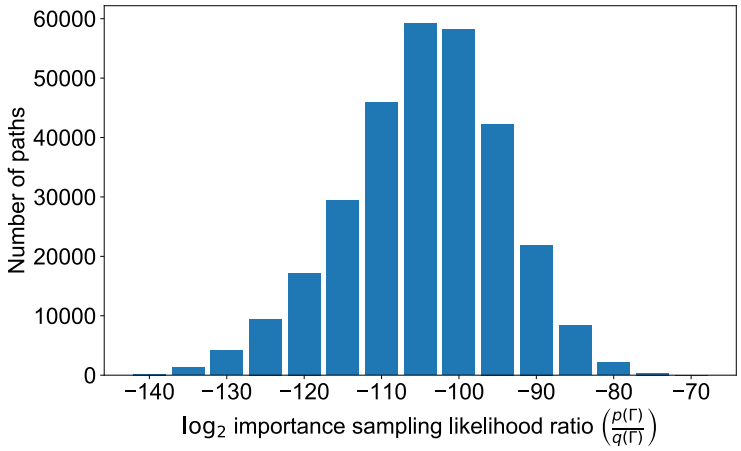


Figure 2.2: Importance sampling likelihood ratio distribution of 300K inheritance paths, simulated from a single patient panel within the BALSAC genealogy.

If we have information about the genotype of close relatives to carriers, we simply simulate the transmission of these known major alleles, forbidding coalescence between lineages carrying different alleles. Because we do not assume a common origin within the genealogy for major alleles, their inheritance can be simulated without importance sampling to ensure coalescence.

Carriers of major alleles who are not closely related to cases have a weak individual impact on trajectory likelihoods, but collectively can contribute substantially. Rather than simulating allele climbing for millions of major alleles (which would be feasible but slow), we treat unrelated homozygotes for the major allele in an average manner. In addition to being numerically convenient, this approach is the best we can do when population-wide allele prevalence were estimated from a sample without genealogical information, as is the case for the CAID allele examined in the Applications section.

We use a 'climb-then-drop' approach, climbing from the minor carriers to generate inheritance paths, then dropping alleles from individuals within simulated inheritance paths back down to the present-day population to estimate major and minor allele prevalence in the general population. This climb-then-drop approach is possible because of the fixed genealogy: a full simulation of the transmission of alleles through a genealogy requires choosing a paternal or maternal transmission at each node, but the order in which these choices are made does not affect the likelihood. We can therefore first simulate the transmissions among ancestors to the known carriers, by climbing alleles and ensuring that they find a common ancestor, and only then proceed to assign the downstream transmissions by dropping these simulated alleles through the

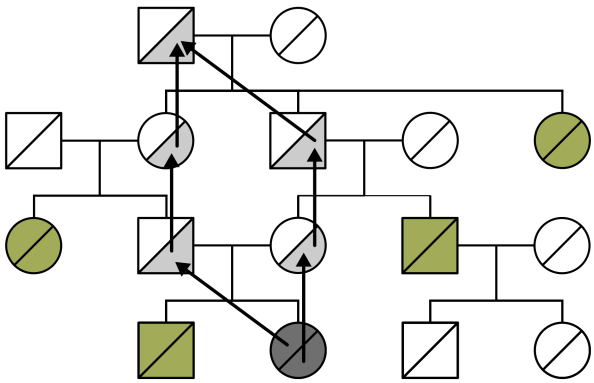


Figure 2.3: The *boundary* of an inheritance path is the set of first-generation descendants (in green) of any individuals within the path (in gray).

rest of the genealogy.

Let F be the random variable representing the minor allele frequency in the present-day population and f its observed value in a population sample collected independently of the genealogy. Dropping alleles from transmission history Γ allows us to estimate $P(F = f|\Gamma, S, A = a)$, the distribution of the allele frequency conditional on Γ and the observed event S (see Appendix 7.1.3 for mathematical details). Appendix 7.1.2 shows that we can estimate the joint probability of the observed carriers and global allele frequencies as

$$P(S, F = f|A = a) \simeq \frac{1}{M} \sum_{j=1}^{M} \mathbb{1}_a(\Gamma_j) \frac{p(\Gamma_j)}{q(\Gamma_j)} P(F = f|\Gamma_j, S, A = a). \tag{2.5}$$

We can then refine the posterior probability that ancestor a was the origin of the allele within the genealogy by conditioning on F as well as S:

$$P(A = a|S, F = f) = \frac{P(S, F = f|A = a)}{\sum_{a' \in \mathcal{A}} P(S, F = f|A = a')}. \tag{2.6}$$

Directly estimating $P(F = f|\Gamma_j, S, A = a)$ by dropping alleles from Γ_j is possible but computationally costly: to get a distribution of f, we need many dropping simulations for each Γ_j. To avoid this computational cost, we propose an approximation that reuses a single set of dropping simulations across all individuals. A naive approach would estimate the present-day frequency of the minor allele as a sum over dropping contributions from all individuals in Γ_j. Unfortunately, since individuals in Γ_j are parentally related, the contributions of individuals in Γ_j to the present-day allele frequency are necessarily overlapping.

To avoid double-counting, we define the *boundary* $\partial\Gamma_j$ of the inheritance path Γ_j as the offspring of all individuals in the path, excluding those in the path itself (see Fig. 2.3). We then compute the global

allele frequency as a sum over individuals in $\partial\Gamma_j$, assumed to contribute approximately independently to the present-day allele frequency. We validated such estimates of $P(F = f|\Gamma)$ by comparing the results to simulated allele drops from the whole inheritance path, and see excellent agreement (see Fig. 7.3 and Appendix 7.1.3 for mathematical details).

Haplotype sharing

Carriers of the minor allele also share a finite haplotype, and the length of the shared haplotype contains information about its origin and transmission history. As a first step towards incorporating this information, we explicitly model the likelihood of the maximum shared haplotype length - the longest haplotype shared amongst all carriers of the minor allele. A similar derivation can be found in Boehnke et. al. [26]

Since we simulate every transmission event in the genealogy, we can also explicitly model the breakdown of a shared haplotype by recombination. The length of this shared haplotype will be the distance between the first recombination in the 3' direction and the first recombination in the 5' direction.

If we assume that recombination follows a Poisson process with a rate of one recombination per Morgan per generation, the waiting distance until the first recombination in either direction from the locus of interest is exponentially distributed with rate corresponding to the number of transmission events below the most recent common ancestor (MRCA) of the carriers. The distribution of shared haplotype lengths will therefore be a sum of two exponential distributions, or an Erlang 2 distribution. Letting h represent the number of meioses since the MRCA of the carriers, the probability of observing a shared haplotype length L is therefore

$$P(L = l|\Gamma) = \text{Erlang}(2, h).$$

We can then incorporate the probability of observing L into our Monte Carlo estimates, as we did with the global allele frequency in (2.6). The expression for the most likely ancestor becomes

$$P(S, F = f, L = l|A = a) \simeq \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}_a(\Gamma_j)\,\frac{p(\Gamma_j)}{q(\Gamma_j)} P(F = f|\Gamma_j)P(L = l|\Gamma_j). \tag{2.7}$$

We can then refine the posterior probability that ancestor a was the origin of the allele within the genealogy by conditioning on L as well as S and F:

$$P(A = a|S, F = f, L = l) = \frac{P(S, F = f, L = l|A = a)}{\sum_{a' \in \mathcal{A}} P(S, F = f, L = l|A = a')}. \tag{2.8}$$

Regional and Individual Carrier Rate Estimation

Obtaining individual and regional carrier rates is useful for both clinical and public health reasons. In a population such as Quebec with an extensive known genealogy, the known relatedness between individuals can be used to estimate such carrier rates. The posterior probability that individual I carries the minor allele is the proportion of transmission histories for which I is a carrier, among all transmission histories consistent with observations.

We again use importance sampling to simulate ascending histories consistent with the observations, and then descending simulations to estimate the probability that an individual is a carrier, conditional on the ascending genealogy. Appendix 7.1.4 shows shows that we can similarly estimate expected prevalences R_m of the minor allele for arbitrary regions:

$$E[R_m|S, F = f] \simeq \frac{\sum_{j=1}^{M} \frac{p(\Gamma_j)}{q(\Gamma_j)} P(F = f|\Gamma_j) E[R_m|\Gamma_j, F = f, S]}{\sum_{j=1}^{M} \frac{p(\Gamma_j)}{q(\Gamma_j)} P(F = f|\Gamma_j, S)}. \tag{2.9}$$

We compute $E[R_m|\Gamma_j, F = f, S]$ using the 'boundary approximation' described above: R_m is taken to be a sum of independent contributions from individuals in $\partial\Gamma$.

Importance tuning for faster convergence

While the straightforward importance sampling scheme presented above provides a large gain in efficiency compared to unweighted Monte Carlo (on the order of $2^{100} \simeq 10^{30}$ times more efficient), there are natural ways to improve and generalize it further. In this section, we describe a more complex scheme that results in faster convergence. The choice of a scheme only affect the convergence speed of the algorithm, and have no effect on the converged results.

For example, while our scheme guarantees that every simulated inheritance path coalesces within the genealogy, it does not seek to favor maternal or paternal inheritance as long as both have nonzero coalescence likelihood. This is suboptimal when the two choices lead to different coalescence likelihoods.

To encourage alleles of a given type to converge towards each other within the genealogy, we implemented an importance sampling scheme that generates an effective attraction among alleles of the same type by sending messages up and down the genealogy. First, we define $t_k(i, j)$ as the length, in generations, of each genealogical route k connecting individual i with their genealogical ancestor j. The probability of an allele in i having independently been inherited from j is therefore the kinship coefficient

$$P(j \to i) = \sum_k 2^{-t_k(i,j)}. \tag{2.10}$$

Each ancestor in the genealogy then gets a score which is the sum of these probabilities of each observed minor allele copy. An ancestor with a large score is therefore a plausible coalescence point for several carriers.

When choosing a parent to climb to, we want to favor parents with high-scoring ancestors. Specifically, we compute a parental score as the sum of the scores its own ancestor, weighted by kinship coefficient linking the parent to its ancestors. Parents are then sampled proportionately to these weighted scores.

Even though it requires many more computations per iteration, the faster convergence can still lead to much lower computational times. In our simulations and inferences, sampling parents by kinship score reduced the overall compute time by roughly a factor of 4. Comparison of convergence rates are shown in Figs. 7.1 and 7.2: the mean standard deviation of likelihood estimates across all ancestor is reduced by an order of magnitude.

Validation

We first use forward simulations (allele dropping) for validation in the single locus setting. Motivated by the CAID example, we assumed a recessive trait. By dropping alleles through the genealogy from each founder, we generate sets of simulated homozygous patients, as well as an associated allele frequency in the rest of the population. We then evaluate how often the importance sampling method correctly re-identifies the generating founder of each patient panel and whether the posterior probabilities are well-calibrated.

We performed the simulations in the BALSAC Population Register genealogy described above. Because validation of posterior probability calibration is computationally intensive, requiring hundreds of individual inferences, we performed it within a subset of the entire genealogy. This subset had been generated by selecting 140 individuals from the most recent generation and including their complete ascending genealogies up to the founders. The 140 individuals included 12 individuals identified in the CAID study and 128 randomly-selected individuals from the most recent generation (The CAID study membership is not used for this validation step, and all 140 individuals are treated equally in this simulation.) This gave a total of 41,523 individuals in a single genealogy with a maximum depth of 17 generations and a median maximum depth across individuals of 15. We then performed forward simulations, selecting forward simulations for which we had between 5 and 30 homozygous patients, giving 470 simulated patient panels for which we knew the ancestral origin of the shared allele.

We then performed 300K importance sampling climbing simulations on each of these simulated panels. Each simulation estimates posterior probabilities for all common ancestors of the simulated homozygous patients (904 unique founders across all panels). In many cases, only a few ancestors have a high probability and the remaining probabilities are quite low. An example is shown in Fig. 2.4.

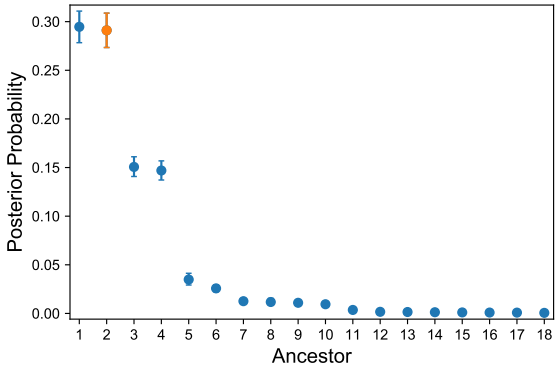


Figure 2.4: Ancestor posterior probabilities for a simulated patient panel. The ancestor generating the panel is shown in orange. Ancestors 1 and 2, as well as 3 and 4, are genealogically indistinguishable founder couples, and are expected to have identical probabilities. Error bars represent uncertainty due to the finite sample size (i.e., the finite number of iterations) in importance sampling. 95% confidence intervals were obtained from bootstrapping over iterations. This source of uncertainty could be further reduced by increasing the number of iterations. Only ancestors with nonzero posterior probability are displayed, and ancestor labels represent ordering by posterior probability for a given simulation. A representative set of simulation results is shown in Figure 7.7.

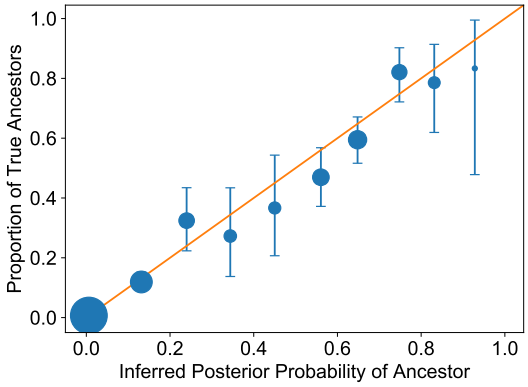


Figure 2.5: Proportion of ancestor clusters that contain the true founding ancestors as a function of cluster posterior probability of containing the true founding ancestor. Error bars represent 95% confidence intervals based on the finite number of observations in each bin. Dot diameter corresponds to the logarithm of this bin count.

Some ancestors are statistically indistinguishable due to symmetries in the genealogy. Monogamous founder couples and grandparent groups connected to the genealogy through a single grandchild are examples. Calculating probabilities for these individuals separately gives no extra information on the likelihood of our simulated inheritance paths, so we sum their probabilities to get a total for the group.

Most ancestors have low posterior probabilities of being the initial carrier. Because we are especially interested in validating posteriors for fairly plausible events, we further group individuals in relatedness clusters, so that we report posterior probabilities that the founder originated in a given relatedness cluster rather than in a given individual (most relatedness clusters are composed of a single founder couple, see Appendix 7.1.5 for details of cluster composition).

The posterior probability of each relatedness cluster, calculated using (2.6), gives an estimate of how often we expect an ancestor from this cluster to be the generating ancestor of that particular patient panel. Fig. 2.5 shows how often a relatedness cluster in a given posterior probability bin contains the true generating ancestor. The means and 95% confidence intervals of this distribution for each bin are obtained under a binomial model (See Appendix 7.1.5 for statistical details).

To validate regional allele frequencies, we used the full BALSAC genealogy. Again performing forward simulations to generate 100 panels of homozygous patients sharing an allele inherited from a single founder, we also recorded the associated allele frequencies in 23 geographic regions of Quebec. We then choose a random sample of 1000 individuals to obtain an estimate f of the global allele frequency. We then use these patient panels S and global allele frequencies f together with (2.9) to compute regional allele frequencies.

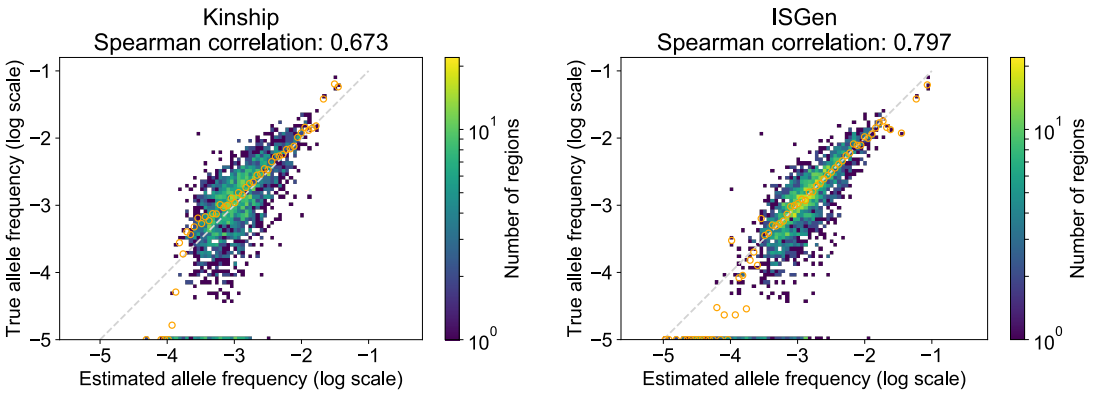


Figure 2.6: Comparison of regional allele frequency estimates based on kinship with known patients and carriers (left column) to those based on inferred allele histories within the full BALSAC genealogical database (right column). We simulated 100 patient panels and corresponding regional allele frequencies. Simulated regional allele frequencies are compared to inference results based on patient panels and estimated global allele frequency. Regions with zero allele frequency in the simulations appear here with frequency 10^{-5}. The asymmetry of the heatmap is due to the logarithmic scale. Orange circles denote the mean true frequency for each estimated frequency bin.

We then compare the inferred results to the true simulated values, shown in Fig. 2.6 and Table 7.3.

We also compare the importance sampling method to a natural alternative, based on kinship scores. When a genealogy is available, pairwise kinship scores give the probability that two individuals are Identical-By-Descent (IBD) at any given locus. Calculating the average kinship of probands in a given region to all known carriers of an allele would give a (potentially biased) estimate of the allele frequency in that region. More details of how we calculated the kinship-based estimates are shown in Appendix 7.1.4, and a comparison of the performance of each method is shown in Fig. 2.6 and Table 7.3. The importance sampling method performed significantly better than the kinship method, with a Spearman correlation of 0.797 with the true allele frequencies, versus 0.673 using kinship.

Application to a Rare Recessive Disease

BALSAC Database and Genotype Data

We apply the importance sampling approach to reconstruct the transmission history and expected distribution of the rare recessive mutation causing Chronic Atrial and Intestinal Dysrhythmia (CAID) in Quebec, Canada, using the population-scale BALSAC genealogy [11]. Constructed from 3 million historical birth,

death, and marriage records, we use here a single fully-connected genealogy of approximately 3.4 million individuals, of which approximately 2.7 million have an associated geographical region. The genealogy has a maximum depth of 17 generations, with most present-day individuals having at least one lineage measuring more than 12 generations. A breakdown of the number of historical records per region is shown in Figure 7.5. Despite its size, the proportion of incorrect links in the BALSAC Quebec genealogies is low, with approximately 1% false paternity [27, 28]. All data was acquired and analyzed in accordance with IRB approval at McGill University under IRB Study No. A01-M48-15A.

In total, 11 patients and four heterozygous carriers of the CAID allele have been identified in Quebec and used in this study, based on genotyping of cases using the Illumina HumanOmni5-Quad chip [29] and on population-based samples as part of the Quebec Regional Population Sample [30]. Of these, all 11 patients and one carrier have been linked to the BALSAC genealogy. The remaining three carriers were collected as part of a global screening effort, during which genealogical information was not obtained. See Appendix 7.1.6 for more details on the screening program.

We assume for this analysis that the minor allele was introduced into the Quebec population by a single European founder. All CAID patients share a 2.9 Mb homozygous segment on chromosome 3, where the causal mutation is located in SGO1 (previously named SGOL1, [MIM: 609168]), with an estimated haplotype age of 30 generations, or 900 years [29]. Because the same CAID mutation was also found in a Swedish patient who shares about 700 Kb with the Quebec 2.9 Mb CAID haplotype, we assume that the mutation was not a *de novo* Quebec mutation [29]. The Genome Aggregation Database [31] gives a present-day frequency of the CAID allele (dbSNP rs199815268) of 0.000237 in Europeans. Thus the single founder assumption, while reasonable, cannot be held with absolute confidence. An approach to extend the present model to multiple founder introductions is outlined in the discussion below. See Appendix 7.1.6 for details on the identification of shared haplotypes among carriers of the CAID allele.

Finally, since CAID is associated with a severe reduction in fecundity, even with modern medical assistance [29], we assume that no homozygote individuals are present in the ascending genealogy, and assign zero likelihood to inheritance histories which contain them.

Estimating the Ascending Allele History

Using ISGen, we then constructed 20 million inheritance paths consistent with the 11 CAID patients and 1 carrier, avoid simulating inheritance paths that do not coalesce to a single ancestor, or which contain ancestral homozygotes for the CAID allele. We calculated the population allele frequency using 3 observed carriers among 900 individuals [32], using (2.7) and (2.8) to integrate this information with the importance

sampling likelihoods.

Among 60 104 distinct ancestors identified in these genealogies, only 31 are founders and common to all CAID carriers. These include 13 founder couples and 5 individual founders who married with non-founders, thus leaving 18 possibly distinguishable genealogical routes for the CAID mutation to enter Quebec.

Two families (given anonymized labels 1 and 2 in Table 2.1) are most likely to have introduced the CAID mutation in the population. Posterior probabilities are shown in Table 2.1, along with confidence intervals from 1000 bootstraps of the simulated inheritance paths and corresponding likelihoods. The combined posterior probability of founder families 1 and 2 is 98.8% (95% confidence interval 0.983-0.991). The two families in total contain 5 founders: family 1 consists of a single monogamous founder couple; family 2 contains a monogamous founder couple with a single child in the genealogy, who forms a monogamous couple with another founder.

Family	Posterior Probability	95% Confidence Interval
1	0.676	(0.599, 0.752)
2	0.312	(0.235, 0.389)
All Others	0.0123	(0.00894 , 0.0171)

Table 2.1: Posterior probabilities of the two families most likely to have introduced the CAID allele into Quebec, along with 95% confidence intervals.

In the case of the CAID allele, the modelling of shared haplotype length has little effect on our estimates of the posterior probabilities of each ancestor, since most common ancestors were at comparable distances in the genealogy. Figure 7.4a shows that the difference between the most-favoured and least-favoured inheritance path is only a factor of 2, and the resulting change to the posterior probabilities of each ancestor by less than 1%, as shown in Fig. 7.4b. A more detailed haplotype sharing analysis may lead to stronger corrections, especially in genealogies with a combination of very recent and older common ancestors.

Fig. 2.7 and Table 7.2 show regional allele frequencies estimated using 1 million simulated inheritance paths, with confidence intervals in Table 7.2 estimated from bootstrapping over inheritance paths. Using the Quebec-wide population frequency estimate of 1/600 for the CAID allele, random mating suggests one affected individual in 360 000 births roughly. However, we find considerable regional heterogeneity, as expected given that the population of Quebec is not genetically homogeneous [33], but formed through a series of regional founder effects [34, 35]. ISGen estimates the CAID allele frequency in Charlevoix to be approximately 1/155, giving a much higher estimated incidence of one affected individual per 24,025 births, assuming random mating.

The full analysis, from simulating inheritance paths to estimating regional prevalences, was performed on a compute cluster in batches of 100K Monte Carlo iterations. Estimating the ascending allele history was

the most computationally costly step, with each batch taking 35 hours to complete on an Intel 3.5GHz Core i7-3770K processor with 16GB of DDR3 RAM. This gives a sizeable total compute time of approximately 280 days, although it is trivial to parallelize.

Regional allele frequencies can be estimated much more efficiently because convergence of estimates is much faster. Estimating regional frequencies took an extra 5 hours per 100K Monte Carlo iterations, giving a total of 40 hours per batch, and 16.6 days for the full 1 million iterations. For those without academic access to such resources, the CAID regional frequency estimates could be completed in a single day on the Google Cloud Platform for CAN$49.58 (40 machines with 2 cores and 7.5GB of memory, 10 hours usage).

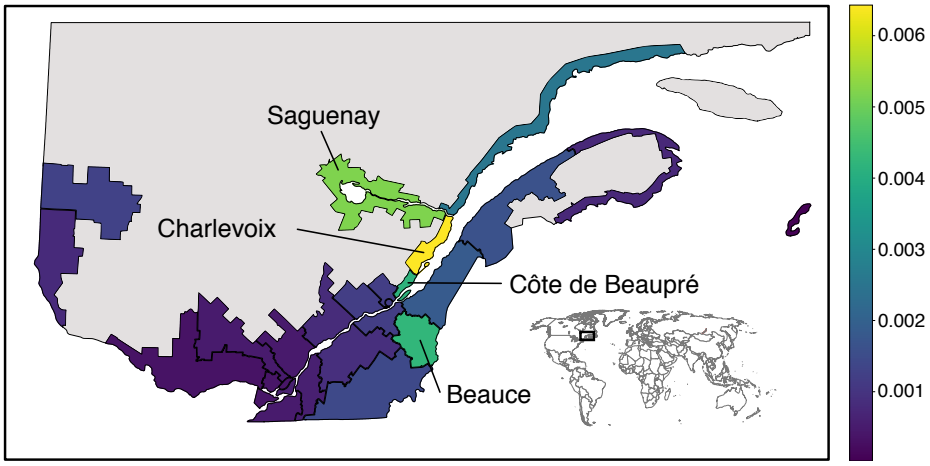


Figure 2.7: Regional expected CAID mutation frequency within the province of Quebec. Grey indicates low-population areas. For fully-labelled regions see Fig. 7.6.

Discussion

Current screening programs do not detect the majority of known rare genetic disorders [36], which cumulatively are estimated to affect up to 2% of couples [37]. Screening programs for such disorders are already in place in regions where cases are found at relatively higher prevalence [38]. Extending these screening efforts to other regions requires a cost-benefit analysis based on incomplete information: genetic risk remains difficult to assess in regions with small population sizes (where the number of cases is low), or with substantial recent migration.

By identifying regions with high predicted carrier rate, *ISGen* provides useful information for the most efficient extension of screening programs. Where genealogies are available, the importance sampling scheme presented here represents a simple way to estimate regional carrier rates, without going through the time- and resource-consuming process of recruiting and genotyping individuals in each region. For example, *ISGen*

predicts the highest allele frequency in Quebec for the CAID mutation at 0.64% in the Charlevoix region, even though no cases or carriers have been reported in that area. This is 24% more than in the more populated Saguenay region where most cases have been identified and screening programs are already in place.

The model considered still has limitations. For example it assumes that the genealogy is specified exactly. However, in some cases, the model defined by Eqs. (2.5) or (2.7) can be sensitive to genealogical errors. Allowing for adoption or false paternity is conceptually straightforward, but there are enough statistical and computational subtleties that we will leave this for future work. In short, even though it is straightforward to allow for adoption, missed paternities, or incorrect genealogical links while simulating inheritance histories, the importance sampling scheme that we have used above must be modified, as any ancestor now has a small but nonzero probability of contributing the minor allele. The same argument holds for multiple founding ancestors: it is straightforward to allow for multiple ancestors to have contributed an allele (this would happen naturally if we did not use importance sampling!), but allowing for multiple founders while ensuring rapid convergence requires more careful tuning of the importance sampling scheme.

We presented and implemented *ISGen* for neutral and lethal recessive alleles because the simple relationship between carrier fitness and genealogical structure simplifies the formulation and implementation. We leave for future work the analysis of alleles with more general modes of inheritance and fitness effects. In particular, estimates of fitness have been performed within the BALSAC genealogy using the effective family size, or number of married children [12]. Family sizes can be influenced by geographic and cultural factors as well as by selection, and their modelling requires more careful discussion.

More generally, we have shown that inferring population-scale allele transmission histories is computationally feasible, even in genealogies containing millions of individuals. We have also made the corresponding software package *ISGen* open-source and freely available at the URL indicated below. Understanding the relative roles of drift and selection in shaping the distribution of disease variants has applications for both medical and evolutionary genetics. Demographic events such as serial founder effects, range expansions, and assortative mating can dramatically alter variant distributions and the effect of natural selection [34, 12]. The increasing availability of large-scale genealogical data, together with statistical tools to infer allele transmissions over time, provides an opportunity to study autosomal inheritance with a unprecedented level of detail.

Acknowledgements

The authors wish to thank M.-H. Roy-Gagnon for her contributions in the early stages of this project, and S. Girard and E. Thompson for useful discussions. This research was undertaken, in part, thanks to funding from the Canada Research Chairs program, the Alfred P. Sloan Foundation, CIHR Discovery grant MOP-136855, FQRNT scholarship 209362, and the FRQS-funded Réseau de Médecine Génétique Appliquée.

Declaration of Interest

The authors declare no conflict of interest.

Web Resources

ISGen, `https://github.com/DomNelson/ISGen`

BALSAC Project, `http://balsac.uqac.ca/`

gnomAD Browser, `http://gnomad.broadinstitute.org/`

OMIM, `http://www.omim.org/`

Quebec Reference Sample, `http://www.quebecgenpop.ca/`

Bibliography

[1] McKusick-Nathans Institute of Genetic Medicine. (2018). *Online Mendelian Inheritance in Man, OMIM.* `http://www.omim.org/`. Accessed May 14, 2018. 2018.

[2] M. H.D. Larmuseau et al. "Genetic genealogy comes of age: Perspectives on the use of deep-rooted pedigrees in human population genetics". In: *American Journal of Physical Anthropology* 150.4 (2013), pp. 505–511. ISSN: 00029483. DOI: `10.1002/ajpa.22233`.

[3] Vigdis Stefansdottir et al. "The use of genealogy databases for risk assessment in genetic health service: A systematic review". In: *Journal of Community Genetics* 4.1 (2013), pp. 1–7. ISSN: 1868310X. DOI: `10.1007/s12687-012-0103-3`.

[4] Andrejs. Plakans, Tamara K Hareven, and Muse. *Family History at the Crossroads: A "Journal of Family History" Reader*. Princeton, N.J.: Princeton University Press, 1987. ISBN: 9781400886913 1400886910.

[5] R D Macmillan. "Screening women with a family history of breast cancer–results from the British Familial Breast Cancer Group." In: *European journal of surgical oncology* 26.2 (2000), pp. 149–52. ISSN: 0748-7983. DOI: `10.1053/ejso.1999.0759`.

[6] Hélène Vézina et al. "Molecular and genealogical characterization of the R1443X BRCA1 mutation in high-risk French-Canadian breast/ovarian cancer families". In: *Human Genetics* 117.2-3 (2005), pp. 119–132. ISSN: 03406717. DOI: `10.1007/s00439-005-1297-9`.

[7] HD Nelson et al. "Genetic risk assessment and BRCA mutation testing for breast and ovarian cancer susceptibility: systematic evidence review for the U.S. Preventive Services Task Force." In: *Annals of internal medicine* 143.5 (2005), pp. 362–379. ISSN: 0003-4819.

[8] American Gastroenterological Association. "American Gastroenterological Association medical position statement: hereditary colorectal cancer and genetic testing." In: *Gastroenterology* 121.1 (2001), pp. 195–197. ISSN: 0016-5085.

[9] Paula W Yoon et al. "Can family history be used as a tool for public health and preventive medicine?" In: *Genetics in Medicine* 4.4 (2002), p. 304.

[10] Steven C Hunt, Roger R Williams, and Gary K Barlow. "A comparison of positive family history definitions for defining risk of future disease". In: *Journal of chronic diseases* 39.10 (1986), pp. 809–821.

[11] BALSAC. *BALSAC Population Database: 2016-2017 Annual Report.* http://balsac.uqac.ca/english/files/2018/01/BALSAC_RA2017_EN_page_WEB_v2-1.pdf. Accessed April 8, 2018. 2018.

[12] C. Moreau et al. "Deep Human Genealogies Reveal a Selective Advantage to Be on an Expanding Wave Front". In: *Science* 334.6059 (2011), pp. 1148–1150. ISSN: 0036-8075. DOI: 10.1126/science.1212880.

[13] Héloïse Gauvin et al. "GENLIB: An R package for the analysis of genealogical data". In: *BMC Bioinformatics* 16.1 (2015). ISSN: 14712105. DOI: 10.1186/s12859-015-0581-5.

[14] Jessica X Chong et al. "A population-based study of autosomal-recessive disease-causing mutations in a founder population". In: *American Journal of Human Genetics* 91.4 (2012), pp. 608–620. ISSN: 00029297. DOI: 10.1016/j.ajhg.2012.08.007.

[15] Charles Y K Cheung, Elizabeth A. Thompson, and Ellen M. Wijsman. "GIGI: An approach to effective imputation of dense genotypes on large pedigrees". In: *American Journal of Human Genetics* 92.4 (2013), pp. 504–516. ISSN: 00029297. DOI: 10.1016/j.ajhg.2013.02.011.

[16] Alan Medlar et al. "SwiftLink: Parallel MCMC linkage analysis using multicore CPU and GPU". In: *Bioinformatics* 29.4 (2013), pp. 413–419. ISSN: 13674803. DOI: 10.1093/bioinformatics/bts704.

[17] Adam P Levine et al. "Genetic Complexity of Crohn's Disease in Two Large Ashkenazi Jewish Families". In: *Gastroenterology* 151.4 (2016), pp. 698–709. ISSN: 15280012. DOI: 10.1053/j.gastro.2016.06.040.

[18] Charles Y.K. Cheung, Elizabeth Marchani Blue, and Ellen M Wijsman. "A statistical framework to guide sequencing choices in pedigrees". In: *American Journal of Human Genetics* 94.2 (2014), pp. 257–267. ISSN: 00029297. DOI: 10.1016/j.ajhg.2014.01.005.

[19] Oren E. Livne et al. "PRIMAL: Fast and Accurate Pedigree-based Imputation from Sequence Data in a Founder Population". In: *PLoS Computational Biology* 11.3 (2015), pp. 1–14. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1004139.

[20] Eric Sobel, Haydar Sengul, and Daniel E Weeks. "Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees". In: *Human Heredity* 52.3 (2001), pp. 121–131. ISSN: 00015652. DOI: 10.1159/000053366.

[21] Simon C Heath. "Markov Chain Monte Carlo Segregation and Linkage Analysis for Oligogenic Models". In: *American Journal of Human Genetics* 61 (1997), pp. 748–760.

[22] Charles J Geyer and Elizabeth A Thompson. "Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference". In: *Journal of the American statistical association* 90.431 (1995), pp. 909–920.

[23] Philip J. Lupo et al. "Family history of cancer and childhood rhabdomyosarcoma: A report from the children's oncology group and the Utah Population Database". In: *Cancer Medicine* 4.5 (2015), pp. 781–790. ISSN: 20457634. DOI: 10.1002/cam4.448.

[24] Daniel F. Gudbjartsson et al. "Sequence variants from whole genome sequencing a large group of Icelanders". In: *Scientific Data* 2 (2015), pp. 1–11. ISSN: 20524463. DOI: 10.1038/sdata.2015.11.

[25] Joanna Kaplanis et al. "Quantitative analysis of population-scale family trees with millions of relatives". In: *Science* 360.April (2018), pp. 171–175.

[26] Michael Boehnke. "Limits of Resolution of Genetic Linkage Studies: Implications for the Positional Cloning of Human Disease Genes". In: *American Journal of Human Genetics* 55 (1994), pp. 379–390.

[27] Evelyne Heyer et al. "Estimating Y Chromosome Specific Microsatellite Mutation Frequencies using Deep Rooting Pedigrees". In: *Human Molecular Genetics* 6.5 (1997), pp. 799–803. ISSN: 0964-6906.

[28] E Heyer et al. "Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees." In: *American Journal of Human Genetics* 69.5 (2001), pp. 1113–1126. ISSN: 0002-9297.

[29] Philippe Chetaille et al. "Mutations in SGOL1 cause a novel cohesinopathy affecting heart and gut rhythm". In: *Nature Genetics* 46.11 (2014), pp. 1245–1248. ISSN: 15461718. DOI: 10.1038/ng.3113.

[30] Quebec Reference Sample. (2010). *Quebec reference sample: Population genetics and genetic epidemiology in Quebec*. http://www.quebecgenpop.ca/. Accessed May 14, 2018. 2010.

[31] Monkol Lek et al. "Analysis of protein-coding genetic variation in 60,706 humans". In: *Nature* 536.7616 (2016), pp. 285–291. ISSN: 14764687. DOI: 10.1038/nature19057. eprint: 030338.

[32] Philip Awadalla et al. "Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics". In: *International Journal of Epidemiology* 42.5 (2013), pp. 1285–1299. ISSN: 03005771. DOI: 10.1093/ije/dys160.

[33] CR Scriver. "Human genetics: lessons from Quebec populations." In: *Annual review of genomics and human genetics* 2 (2001), pp. 69–101. ISSN: 1527-8204.

[34] Claude Bhérer et al. "Admixed ancestry and stratification of Quebec regional populations". In: *American Journal of Physical Anthropology* 144.3 (2011), pp. 432–441. ISSN: 0002-9483.

[35] M Labuda et al. "Linkage disequilibrium analysis in young populations: pseudo-vitamin D-deficiency rickets and the founder effect in French Canadians." In: *American Journal of Human Genetics* 59.3 (1996), pp. 633–643. ISSN: 0002-9297.

[36] Lidewij Henneman et al. “Responsible implementation of expanded carrier screening”. In: *European Journal of Human Genetics* 24.6 (2016), e1–e12. ISSN: 14765438. DOI: 10.1038/ejhg.2015.271.

[37] Hans-Hilger Ropers. “On the future of genetic risk assessment”. In: *Journal of Community Genetics* 3.3 (2012), pp. 229–236. ISSN: 1868-310X. DOI: 10.1007/s12687-012-0092-2.

[38] Jessica Tardif, Annabelle Pratte, and Anne-Marie Laberge. “Experience of carrier couples identified through a population-based carrier screening pilot program for four founder autosomal recessive diseases in Saguenay-Lac-Saint-Jean”. In: *Prenatal Diagnosis* 38 (2017), pp. 27–74. ISSN: 10970223. DOI: 10.1002/pd.5055.

Preface to Chapter 3

As we have seen with ISGen, inferring genealogies for even single alleles has practical applications for genetic screening of rare diseases. Performing such inference along the whole genome potentially offers substantially more value, with applications ranging from demographic inference, mutation-rate estimation, detection of introgression, and identification of genomic regions under selection, as well as more efficient storage requirements for biobank-size datasets [54, 56, 59]. However it is not currently possible to infer genealogies for hundreds of thousands of whole genomes in a fully Bayesian framework, requiring the use of simpler heuristic methods. This is justified by the sheer volume of data which can be integrated, but also suggests that careful validation is especially important in order to understand the uncertainty and possible biases of the inferred genealogies.

Simulations are a natural validation method, but face challenges in matching the scale of modern biobanks. The most efficient simulators, based on Hudson's coalescent theory [61, 58], struggle to maintain realistic relatedness among samples across long regions, as we show in the following manuscript. This efficiency however cannot be sacrificed, as performing simulations at this scale is computationally challenging. We explore now an extension of the highly-efficient `msprime` simulation software [58] allowing simulations to be performed under the Wright-Fisher model, vastly improving relatedness among simulated individuals while further increasing computational efficiency at whole-genome scale.

Chapter 3

Accounting for long-range correlations in genome-wide simulations of large cohorts

Dominic Nelson[1], Jerome Kelleher[2], Aaron P. Ragsdale[1], Claudia Moreau[3], Gil McVean[2], Simon Gravel[1]

Published in *PLOS Genetics* 16(5): e1008619 on May 5th, 2020.

1. McGill University and Genome Québec Innovation Centre, McGill University, Montréal, Québec, Canada

2. Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, United Kingdom,

3. Centre Intersectoriel en Santé Durable, Université du Québec à Chicoutimi, Saguenay, Québec, Canada

Abstract

Coalescent simulations are widely used to examine the effects of evolution and demographic history on the genetic makeup of populations. Thanks to recent progress in algorithms and data structures, simulators such as the widely-used `msprime` now provide genome-wide simulations for millions of individuals. However, this software relies on classic coalescent theory and its assumptions that sample sizes are small and that the region being simulated is short. Here we show that coalescent simulations of long regions of the genome exhibit large biases in identity-by-descent (IBD), long-range linkage disequilibrium (LD), and ancestry patterns, particularly when the sample size is large. We present a Wright-Fisher extension to `msprime`, and show that it produces more realistic distributions of IBD, LD, and ancestry proportions, while also addressing more subtle biases of the coalescent. Further, these extensions are more computationally efficient than state-of-the-art coalescent simulations when simulating long regions, including whole-genome data. For shorter regions, efficiency can be maintained via a hybrid model which simulates the recent past under the Wright-Fisher model and uses coalescent simulations in the distant past.

Author summary

Coalescent theory has provided deep theoretical insight into patterns of human diversity. Implementations of coalescent models in simulation software such as `ms` have further provided tools to interpret thousands of genomic studies. Recent technical progress has allowed for a dramatic increase in the scale at which genomes can be both measured and simulated, opening up opportunities for a finer understanding of evolutionary biology. However, we show that coalescent simulations of long regions of the genome exhibit large biases in sample relatedness, distorting haplotype sharing and ancestry patterns in simulated cohorts. We trace these biases to basic assumptions of the coalescent model, and show how the assumptions can be relaxed to provide a better description of the observed patterns of genetic polymorphism at a fraction of the computational cost.

Introduction

Simulations of genome evolution are widely used in the development of computational tools for statistical and population genetics research (e.g., [1, 2, 3, 4, 5, 6]). Coalescent theory has been used extensively for this purpose, with Hudson's `ms` simulation program [7] having been cited over two thousand times since its publication in 2002. The more recent `msprime` coalescent simulation software [8] implements Hudson's

original algorithm [9], but with a performance increase of several orders of magnitude. This is achieved largely through the introduction of a new data structure, the succinct tree sequence [10, 11], which is extremely efficient at storing genetic variation. For example, simulating a 100 megabase region in a sample of 100,000 individuals generates an 88MB uncompressed succinct tree sequence, whereas the Newick tree format used by `ms` takes approximately 3.5TB of space [8].

Simulated data are useful to the extent that they accurately reflect real genetic variation. However, the coalescent is known to be biased relative to the Wright-Fisher model when the sample size is large [12] or for events in the recent past [13]. However, these biases have had limited practical impact because collecting such large empirical data sets was prohibitively costly and the simulation of such large samples was computationally overwhelming. Both limitations have now been lifted: sequencing datasets now regularly include thousands of sequenced genomes, and `msprime` can simulate hundreds of thousands of genomes on a laptop computer. The assumptions of the underlying coalescent models should be carefully reexamined in this context.

We highlight qualitative and quantitative inaccuracies in coalescent simulations of long regions, due to violated assumptions of the underlying genealogical model. We implement an extension to `msprime` which corrects the majority of these biases via a backwards-in-time Wright-Fisher model within `msprime` (see overview in Methods section and S1 Appendix), which generates biologically plausible genealogies regardless of sample size (a separate implementation of such a model, without using succinct tree sequences, can also be found in [14]). Our backwards-in-time Wright-Fisher simulations are also much faster than coalescent simulations for large samples and long regions. For shorter regions, the coalescent is slightly faster. Using a hybrid approach with Wright-Fisher dynamics in the recent past and coalescent dynamics further back in time (as was done in [13]) preserves the computational advantages of the coalescent with the long-range accuracy of the Wright-Fisher model for shorter genomic regions.

Motivation

This work was motivated by our observation that large-scale coalescent simulations resulted in unrealistic relatedness among samples, where nearly every pair of simulated individuals were second- or third-degree cousins according to the time to their most recent common ancestor. This is because individuals had too many simulated ancestors: whereas diploid individuals carry at most 2^t ancestors at generation t in the past, coalescent simulations allow for many more ancestors.

This excess of ancestors is a side effect of how Hudson's coalescent algorithm models recombination. Hudson's coalescent model assumes a small region being simulated [15], and so does not account for multiple

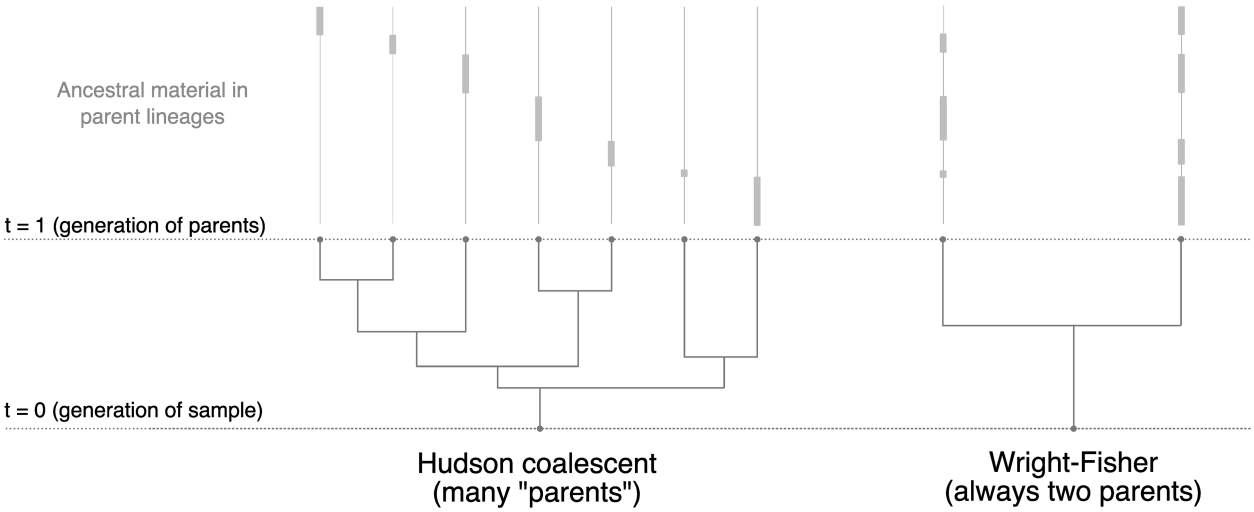


Figure 3.1: Comparing coalescent and Wright-Fisher lineages one generation in the past. A schematic of simulated lineages for a haploid sample with a single long chromosome. In the coalescent, each recombination event creates a new, independent lineage, leading to an unrealistic number of simulated parents. The Wright-Fisher model allows for back-and-forth recombination, so recombination events alternately assign genetic material between only two parental lineages. Multiple chromosomes exaggerate the difference, segregating as expected in the Wright-Fisher model but adding extra lineages under the coalescent.

simultaneous recombinations during meiosis. The per-generation recombination rate in long genomic regions is maintained by multiple recombinations occurring at different times, with each recombination introducing a new ancestral lineage. This can lead to more than two ancestors within one generation (Fig 3.1).

This property of the coalescent recombination model is often innocuous when regions simulated are too short for back-and-forth recombinations to occur, or when the number of lineages is small enough that long range correlations are practically negligible [13, 16]. In larger samples, or under migration models, recent events induce long-range correlations along the genome [12, 17, 18, 19]. For example, individuals with a recent migrant ancestor are likely to have migrant ancestry in several chromosomes, and this is not accounted for by Hudson's coalescent. Significant differences have further been observed between the simulated genealogies of coalescent and Wright-Fisher models at a single locus [13, 14], such as the more rapid decay in the number of lineages over time in the Wright-Fisher model when sample size is large. Model differences become even more pronounced over long regions, where correlations between distant gene genealogies must be taken into account.

To highlight the magnitude of the genealogical distortions which can occur, we first use both the coalescent and Wright-Fisher models to simulate haploid sample sizes from 500 to 10,000 in a diploid population with size 10,000 and growth rate 0.001. Each sample contains 22 chromosomes of realistic lengths. Fig 3.2 shows

that for 10,000 samples the number of lineages in the coalescent simulation increases very rapidly to reach 10 times the haploid population size $2N$ (This issue was also raised in [20, 21]). Simulations with smaller sample sizes also show a rapid growth in number of lineages to beyond the haploid population size, but the growth is slower and the excess is less pronounced than in larger samples. In the Wright-Fisher simulation, the initial growth in number of lineages is much slower and can never exceed the haploid population size, regardless of sample size.

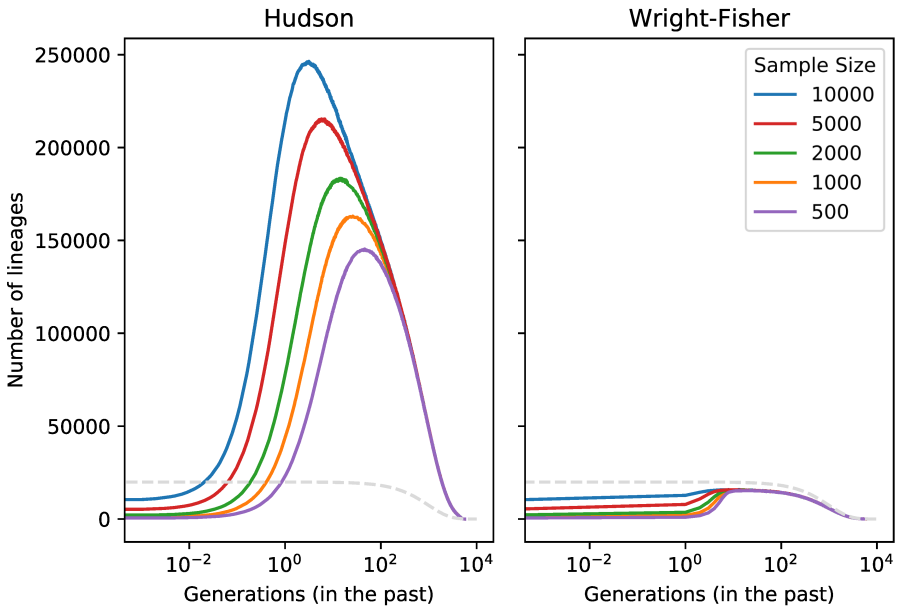


Figure 3.2: Number of surviving lineages over time in coalescent and backwards-in-time Wright-Fisher dynamics. We simulated a varying number of haploid whole genomes with 22 chromosomes of realistic lengths in a population of 10,000 diploid individuals. Dotted line shows effective population size. The implementation for simulations with multiple chromosomes is described in S1 Appendix.

While genealogical distortions are most clear in the first few generations, this explosion of lineages also affects genealogies in the more distant past. Fig 3.2 also shows that, despite rapid coalescence lowering the initial spike in the number of lineages, their number remains above the population size for hundreds of generations into the past. The effect is even more dramatic within a constant-sized population, with S2 Figure showing a case where the number of lineages remains above the effective population size for more than 100,000 generations in the past.

The number of lineages cannot be observed directly from genetic data, but these genealogical distortions have consequences for commonly used measures of genetic diversity.

Results

In this section, we first highlight qualitative differences in multi-locus statistics between the coalescent and backwards Wright-Fisher models, and we show that the Wright-Fisher models provide a better description of the data while increasing tractability.

Distribution of IBD

Under the Wright-Fisher model, diploid inheritance constrains the possible gene genealogies [12] and introduces correlations in IBD sharing along long simulated regions: two samples with a recent common ancestor may be IBD at several distant positions of their genome (for example on different chromosomes). In the coalescent, gene genealogies of unlinked loci are constructed independently, and do not capture this effect [12].

Modelling relatedness patterns is important in large cohorts, where cryptic relatives are common [22, 23]. To illustrate the significance of explicitly modelling diploid inheritance in a sample with close relatives, we compared simulated cohorts to genotype data from participants of the Genizon Biobank containing 8,435 individuals from the province of Quebec, Canada [24]. A description of this biobank and IBD detection methods is given in S4 Appendix. Pairwise IBD patterns observed in this cohort are shown in Fig. 3.3.

We simulated 5,000 human haploid whole genomes (chromosome lengths and recombination rates are described in S1 Appendix) in a diploid population of constant size 10,000 under the coalescent and Wright-Fisher models, and used the simulated genealogies to extract IBD segments inherited from common ancestors up to 5 generations in the past. Closer relatedness means more IBD segments and longer average length, leading to a relationship between number of segments and total length of IBD which is typically used in identifying relative status [22]. Since the detection of very short IBD segments is challenging in practice, we counted only simulated IBD segments greater than 5 centimorgans, in both simulations and the data.

Fig 3.3 shows the difference between the two models, with the Wright-Fisher model showing excellent qualitative agreement with the Genizon data. Quantitative differences are expected since simulations were performed in a non-monogamous randomly-mating population. By contrast, the coalescent model exhibits far too few IBD segments for closely related individuals and poor clustering by TMRCA. An analytical model for the expected number and length of shared ancestry segments (shown as white dots in Fig 3.3) is provided in S3 Appendix. The separated cluster predicted by the Wright-Fisher model represents simulated half-siblings: neither full- nor half-siblings are present in the Genizon data. Other relationships also form clusters that overlap due to variance in amounts of genetic material shared IBD. Residual differences between

Wright-Fisher simulations and theoretical predictions in Fig 3.3 have to do with the requirement that IBD segments be at least 5cM to be detected. Better agreement could be achieved by using a cutoff of 1cM in simulations (see S3 Figure).

The distribution of long IBD segments between related individuals is primarily determined by their degree of recent relatedness. For example, even though the population history and sampling process affects the *number* of sampled first cousins, the recent IBD relatedness *among* first cousins in large outbred populations is relatively independent of history and sampling: This is why the simulated and empirical distributions observed on Fig 3.3 are in good agreement despite differences in population sizes, and why the theoretical predictions that describe both are independent of the population demography. Because the number of close relatives changes with sampling and population size, the discrepancy between coalescent and Wright-Fisher models is more acute for large sample sizes (see S3 Figure and S4 Figure for simulations under different models). Yet S3 Figure shows clear differences between Wright-Fisher and coalescent models with $N_e = 10,000$ and 500 samples. More generally, Shchur et. al. (2018) [23] calculated the expected number of p-th cousins in a sample of size K taken from a population of effective size N. In a monogamous Wright-Fisher population, when $K/N = 0.2$, we expect approximately 55% of samples to have a first cousin, and 95% to have a second cousin within the cohort.

The long-range correlations induced by genealogical relatedness can also be measured as linkage disequilibrium between distant loci. This LD is used to estimate sizes of small populations in conservation genetics [25, 26]. Hudson's coalescent does not capture such LD patterns [17], whereas the Wright-Fisher extension to `msprime` predicts the patterns of LD expected under diploid mating (see S2 Appendix).

Ancestry variance following admixture

In admixed populations, simulations should capture patterns of ancestry variation among present-day samples. The distribution of ancestry within recently admixed populations can be strongly dependent on pedigree structure [18], making coalescent simulations of these scenarios problematic.

We consider the variance of ancestry proportions following a single pulse of migration. Ancestry variance can be divided into genealogical variance and recombination variance [27]. In the first few generations after admixture, variance is driven by genealogical differences in the number of migrant ancestors of each individual. As time goes on, each present-day individual has more ancestors from the admixed generation, exponentially reducing this source of variance. After roughly 10 generations, variation in the amount of genetic material received from each migrant ancestor becomes a stronger source of variance [27].

We performed whole-genome simulations to evaluate how well the Wright-Fisher and coalescent models

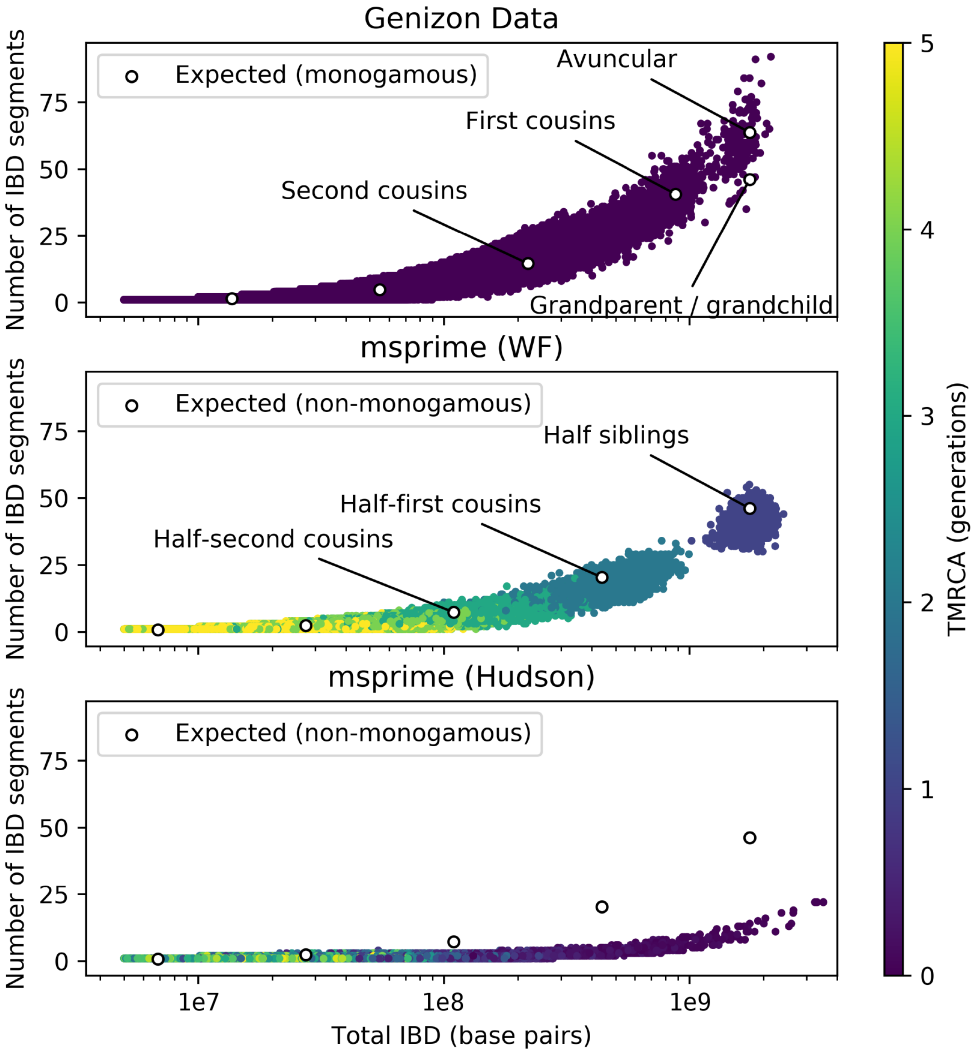


Figure 3.3: Number of IBD segments between pairs of individuals versus total length of shared IBD segments. 22 chromosomes of realistic lengths, simulated under Wright-Fisher model (middle) and coalescent (bottom), compared to data from 8,435 individuals from the Genizon Biobank (top), as well as the analytical expectation under Eqs (1), (2), (3), and (4) in S3 Appendix (white circles). Siblings were filtered from the Genizon cohort, as explained in S4 Appendix. Simulations contained 5,000 haploid samples with a diploid population size of 10,000. The isolated cluster in the Wright-Fisher simulations reflects the discrete nature of possible genealogical relationships (siblings, cousins, etc.) in the Wright-Fisher model.

capture variance in ancestry. Fig 3.4 shows ancestry variance from simulations of 80 haploid samples in a diploid population of size 80, and a single event of 30% admixture at varying time in the past. These parameters were chosen to match those in [27], but here again the qualitative patterns depend weakly on the sample size and older demographic history. The approximate expected values are derived from an argument

similar to the one presented in the supplement for IBD sharing and outlined in [27].

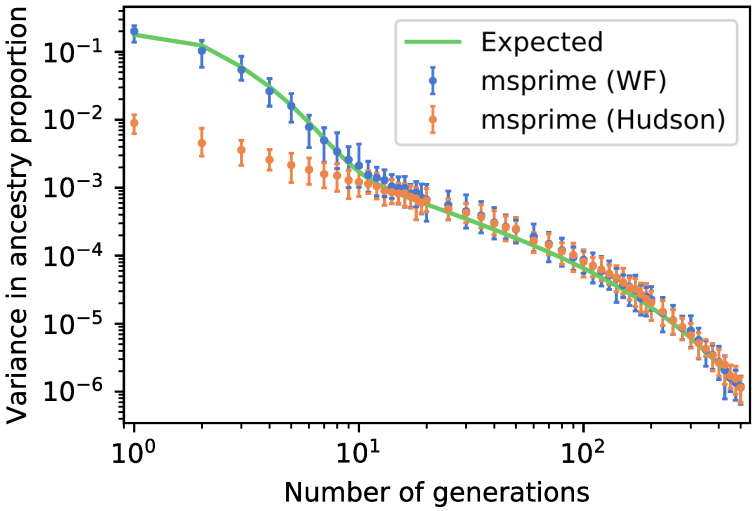


Figure 3.4: Variance in ancestry after a single admixture event, as a function of time since admixture. Calculated from 80 haploid samples in a diploid population of size 80, with 30% admixture proportions. Error bars show 95% confidence intervals over 50 simulations.

The Wright-Fisher model captures both short- and long-term variance in ancestry, as expected. In the coalescent simulations the initial phase of genealogical variance is not present, leading to a 20-fold underestimate of the variance in ancestry. Lacking a diploid population pedigree, whole-genome coalescent simulations of recently admixed populations do not reflect the distribution of ancestry expected in a large cohort, even under an idealized random-mating scenario.

Other genealogical effects

Bhaskar et al. [13] showed that simultaneous coalescences in the Wright-Fisher model lead to more singletons and fewer doubletons than in the coalescent, which was verified in [14]. S1 Figure and S1 Table replicate these single-locus results. King et al. [17] pointed out correlation patterns among unlinked loci induced by genealogical relatedness – these results correspond to the infinite-recombination distance in S2 Appendix.

Performance

The main advantage of `msprime` over alternate simulators is speed and scalability. This is achieved by efficient algorithms and, especially, new data structures for storing and manipulating ancestral states throughout a simulation. We therefore need to ensure that the present modification preserves these advantages.

Hudson's coalescent algorithm avoids simulating recombination and coalescent events that do not affect genetic variation in the present sample. Whereas our Wright-Fisher implementation must iterate over all discrete generations, Hudson's coalescent can traverse long stretches of time in a single step if there are no such events. The Hudson model is therefore more efficient than the Wright-Fisher model when the number of lineages is small, as can happen in small samples and short genomic regions, or in the distant past. However, Fig 3.2 shows that the number of lineages in whole-genome coalescent simulations is so high that the time between events is on average much less than a single generation. Furthermore, these lineages come at an additional memory and computational cost for the coalescent model. This naturally suggests using a hybrid approach with Wright-Fisher dynamics in the recent past and coalescent dynamics in the more distant past, following the approach of Bhaskar et. al. [13].

Our Wright-Fisher extension is integrated with `msprime`'s core simulation framework, and can easily be combined with coalescent simulations as part of a hybrid model. Since the optimal switching time depends on the number of extant lineages and total length of uncoalesced ancestral material, it will vary between different demographic models.

Fig 3.5 shows computation times for Wright-Fisher, Hudson coalescent, and hybrid simulations of 1,000 haploid samples within a population of constant size 10,000. The pure Wright-Fisher simulations are fastest at whole-genome scale, whereas pure coalescent simulations and hybrid approaches are slightly faster for shorter regions. There is a small performance cost to switching models, which explains the slightly longer runtime for the hybrid model with 100 Wright-Fisher generations versus pure coalescent simulations. The hybrid model with 1,000 Wright-Fisher generations compares favourably in terms of performance and accuracy to the coalescent for a wide range of simulated lengths.

Methods

Implementation

To understand the modifications needed to turn msprime into a back-in-time Wright-Fisher simulator, we first outline Hudson's original algorithm to simulate samples under the coalescent model. This brief overview is intended to give context to the modifications which enable Wright-Fisher simulations to be performed in the same framework. More details of how Hudson's algorithm is implemented in `msprime` are given in [8].

First, a number of randomly-mating populations are specified, including effective sizes and migration rates over time. Samples are introduced as haploid lineages within the populations, and the region of the genome being simulated is specified. The algorithm then constructs the genealogy of each locus within this

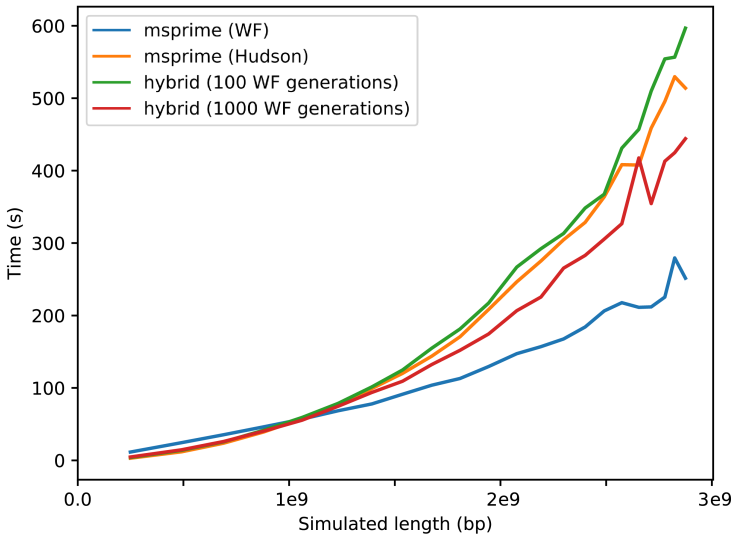


Figure 3.5: Computation time of Hudson coalescent, Wright-Fisher, and hybrid models. Hybrid models used 100 and 1000 Wright-Fisher generations before switching to the coalescent. Simulations contain from 1 to 22 chromosomes of realistic lengths (using the method described in S1 Appendix) in 1,000 haploid samples drawn from a diploid population of constant size 10,000. Results for other population sizes are shown in S5 Figure.

region by tracing its lineages backwards in time and tracking genomic segments that are ancestral to the sample.

To begin, each lineage contains a single ancestral segment spanning the whole simulated genomic region of a sample. As time proceeds backwards, lineages can be split by recombination events (leaving the amount of ancestral material unchanged), or participate in common ancestor events, where any overlapping regions coalesce (reducing the amount of ancestral material). The rate of recombination events depends on the sum of the genetic map distance spanned by ancestral segments carried by all extant lineages, and common ancestor events occur at a rate determined by the number of uncoalesced lineages and the effective population size. Migration events move haploid lineages between randomly-mating populations, and demographic events modify the number of populations or their size and growth rate parameters. Recombination and common ancestor events are generated at rates depending on the amount of extant ancestral material, and the simulation terminates when every position on the genome has a most recent common ancestor

Implementing a back-in-time Wright-Fisher model requires two important changes to Hudson's algorithm. First, rather than drawing a time to the next event from an exponential distribution, we iterate though discrete generations and draw the events which occur at each time. Second, we modify the way recombination events are carried out, to account for the possibility of multiple recombinations in a single transmission: we

model the number and spatial distribution of breakpoints as a Poisson process, with rate equal to the per-generation recombination rate (i.e., the distance in Morgans). This model ensures that each gamete has a unique diploid parent. An overview of this model is illustrated in Fig 3.1 and the detailed order of events occurring at each generation is given in S1 Appendix.

Ethics statement

Access to the Genizon cohort genotyping data was granted under study number A07-M42-15B of the McGill university IRB. Third party data were analysed anonymously so consent was not obtained.

Discussion

While the Wright-Fisher model may generate a more realistic pedigree than the coalescent model in the recent past, it was recognized early on as an idealized model [28, 29]. Our implementation does not track monogamous couples, for example, and therefore will vastly overestimate the prevalence of half-sibs and underestimate full sibs compared to a realistic human cohort. Assortative mating and inbreeding are not accounted for, and the migration model, while biologically plausible, is a simplification of the real migration process (see implementation details in S1 Appendix). Care should be taken in applications which are particularly sensitive to fine-scale mating or migration patterns.

Many of these issues can be addressed by allowing simulations to take place within a pre-specified pedigree, which is a natural extension to our backwards-in-time Wright-Fisher implementation. Rather than drawing genealogical links at random according to demographic parameters, lineages can simply follow a known pedigree. When reaching a pedigree founder, simulations can then continue by reverting to either the Wright-Fisher or the coalescent models. Real pedigrees of any size could then be used, from extended families up to population-scale [30], or they could be generated with the desired patterns of monogamy or assortative mating in a separate step. While conceptually straightforward, maintaining efficiency while simulating within population-scale pedigrees is non-trivial. We leave such an implementation for future work.

Improvements to recombination models is also a natural extension of the present approach. Assigning sexes to parents would allow simulation of the X-chromosome and sex-biased migration. Recombination can be extended to model crossover interference and sex-biased recombination, which have effects on the distribution of IBD [31], as well as non-crossover events.

Finally, the performance of the hybrid model could also be improved. If the number of Wright-Fisher generations were chosen optimally, it is likely to be more efficient than pure Wright-Fisher simulations in nearly all scenarios. Better guidelines for finding this optimal value could be developed, or possibly built

into the simulation framework itself.

The limitations of the coalescent model have been well-studied, but were generally tied to modest effects except in very large cohorts [13]. We have shown significant qualitative and quantitative biases in whole-genome simulations of large, complex cohorts. Analysis of such cohorts is challenging, and simulations are a valuable tool for evaluating disease associations and the effects of demography in this context. We have presented here an extension to `msprime` which corrects major biases and increases performance at whole-genome scale, allowing simulations to continue supporting modern large-scale sequencing efforts.

Bibliography

[1] Christopher S Carlson et al. "Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium". In: *The American Journal of Human Genetics* 74.1 (2004), pp. 106–120. ISSN: 00029297. DOI: 10.1086/381000.

[2] Benjamin F Voight et al. "A map of recent positive selection in the human genome." In: *PLoS biology* 4.3 (2006), e72. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0040072.

[3] Ryan N. Gutenkunst et al. "Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data". In: *PLoS Genetics* 5.10 (2009). ISSN: 15537390. DOI: 10.1371/journal.pgen.1000695.

[4] Heng Li and Richard Durbin. "Inference of human population history from individual whole-genome sequences". In: *Nature* 475.7357 (2011), pp. 493–496. ISSN: 0028-0836. DOI: 10.1038/nature10231.

[5] Na Li and Matthew Stephens. "Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data". In: *Genetics* 165.4 (2003), pp. 2213–2233. ISSN: 00166731. DOI: 10.1534/genetics.104.030692.

[6] Rasmus Nielsen et al. "Genomic scans for selective sweeps using SNP data". In: *Genome Research* 15.11 (2005), pp. 1566–1575. ISSN: 10889051. DOI: 10.1101/gr.4252305.

[7] Richard R Hudson. "Generating samples under a Wright-Fisher neutral model of genetic variation". In: *Bioinformatics* 18.2 (2002), pp. 337–338. ISSN: 13674803. DOI: 10.1093/bioinformatics/18.2.337.

[8] Jerome Kelleher, Alison M. Etheridge, and Gilean McVean. "Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes". In: *PLoS Computational Biology* 12.5 (2016), pp. 1–39. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1004842.

[9] Richard R. Hudson. "Properties of a neutral allele model with intragenic recombination". In: *Theoretical Population Biology* 23.2 (1983), pp. 183–201. ISSN: 10960325. DOI: 10.1016/0040-5809(83)90013-8.

[10] Jerome Kelleher et al. "Efficient pedigree recording for fast population genetics simulation". In: *PLoS computational biology* 14.11 (2018), e1006581. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1006581.

[11] Jerome Kelleher et al. "Inferring whole-genome histories in large population datasets". In: *Nature Genetics* 51.9 (2019), pp. 1330–1338. ISSN: 1061-4036. DOI: 10.1038/s41588-019-0483-y.

[12] John Wakeley et al. "Gene genealogies within a fixed pedigree, and the robustness of kingman's coalescent". In: *Genetics* 190.4 (2012), pp. 1433–1445. ISSN: 00166731. DOI: 10.1534/genetics.111.135574.

[13] Anand Bhaskar, Andrew G Clark, and Yun S Song. "Distortion of genealogical properties when the sample is very large." In: *Proceedings of the National Academy of Sciences of the United States of America* 111.6 (2014), pp. 2385–90. ISSN: 1091-6490. DOI: 10.1073/pnas.1322709111. arXiv: arXiv:1308.0091v1.

[14] Pier Francesco Palamara. "ARGON: fast, whole-genome simulation of the discrete time Wright-fisher process". In: *Bioinformatics* 32.19 (June 2016), pp. 3032–3034. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw355.

[15] Richard R. Hudson. "Gene genealogies and the coalescent process". In: *Futuyma, D. and Antonovics, J. (eds), Oxford Surveys in Evolutionary Biology*. Vol. 7. 1990, pp. 1–44.

[16] Peter R. Wilton et al. "Population structure and coalescence in pedigrees: Comparisons to the structured coalescent and a framework for inference". In: *Theoretical Population Biology* 115 (2017), pp. 1–12. ISSN: 10960325. DOI: 10.1016/j.tpb.2017.01.004.

[17] Léandra King, John Wakeley, and Shai Carmi. "A non-zero variance of Tajima's estimator for two sequences even for infinitely many unlinked loci". In: *Theoretical Population Biology* 122 (2018), pp. 22–29. ISSN: 10960325. DOI: 10.1016/j.tpb.2017.03.002.

[18] Mason Liang and Rasmus Nielsen. "The lengths of admixture tracts". In: *Genetics* 197.3 (2014), pp. 953–967. ISSN: 19432631. DOI: 10.1534/genetics.114.162362.

[19] R. Martin Ball, Joseph E. Neigel, and John C. Avise. "Gene Genealogies within the Organismal Pedigrees of Random-Mating Populations". In: *Evolution* 44.2 (1990), p. 360. ISSN: 00143820. DOI: 10.2307/2409414.

[20] K. J F Verhoeven and Katy L. Simonsen. "Genomic haplotype blocks may not accurately reflect spatial variation in historic recombination intensity". In: *Molecular Biology and Evolution* 22.3 (2005), pp. 735–740. ISSN: 07374038. DOI: 10.1093/molbev/msi058.

[21] Joanna L. Davies et al. "On recombination-induced multiple and simultaneous coalescent events". In: *Genetics* 177.4 (2007), pp. 2151–2160. ISSN: 00166731. DOI: 10.1534/genetics.107.071126.

[22] Brenna M Henn et al. "Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples". In: *PLoS ONE* 7.4 (2012). ISSN: 19326203. DOI: 10.1371/journal.pone.0034267.

[23] Vladimir Shchur and Rasmus Nielsen. "On the number of siblings and p-th cousins in a large population sample". In: *Journal of Mathematical Biology* 77.5 (2018), pp. 1–20. ISSN: 14321416. DOI: 10.1007/s00285-018-1252-8.

[24] Genome Quebec. *Genizon Biobank*. http://www.genomequebec.com/genizon-biobank/. Accessed January 7, 2020. 2020.

[25] Robin S. Waples. "A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci". In: *Conservation Genetics* 7.2 (2006), pp. 167–184. ISSN: 15660621. DOI: 10.1007/s10592-005-9100-y.

[26] Aaron P Ragsdale and Simon Gravel. "Unbiased Estimation of Linkage Disequilibrium from Unphased Data". In: *Molecular Biology and Evolution* (Nov. 2019). ISSN: 0737-4038.

[27] Simon Gravel. "Population genetics models of local ancestry". In: *Genetics* 191.2 (2012), pp. 607–619. DOI: 10.1534/genetics.112.139808.

[28] RA Fisher. *The genetical theory of natural selection.* Clarendon Press, 1930.

[29] Sewall Wright. "Evolution in Mendelian populations". In: *Genetics* 16.2 (1931), p. 97.

[30] BALSAC. *BALSAC Population Database: 2016-2017 Annual Report.* http://balsac.uqac.ca/english/files/2018/01/BALSAC_RA2017_EN_page_WEB_v2-1.pdf. Accessed April 8, 2018. 2018.

[31] Madison Caballero et al. "Surprising impacts of crossover interference and sex-specific genetic maps on identical by descent distributions". In: *bioRxiv* (2019), p. 527655. DOI: 10.1101/527655.

Supporting Information

S1 Appendix Wright-Fisher Implementation Details

S2 Appendix Long-range linkage disequilibrium

S3 Appendix An approximate model for IBD sharing

S4 Appendix The Genizon Biobank

S1 Table Relative difference in mean number of singletons, doubletons, and tripletons under the Wright-Fisher (N_{WF}) and Hudson (N_H) models.

S1 Figure Number of singletons, doubletons, and tripletons simulated under Wright-Fisher and Hudson coalescent models. A 1Mb region was simulated 100 times in 20,000 haploid lineages in a diploid population of 10,000 individuals.

S2 Figure Number of surviving lineages over time in coalescent and back-in-time Wright-Fisher dynamics. We simulated 10,000 haploid whole genomes with 22 chromosomes of realistic lengths in a population of 10,000 diploid individuals. The method for simulating multiple chromosomes is described in S1 Appendix. Similar results were shown in [68].

S3 Figure Number of IBD segments between pairs of individuals versus total length of shared IBD segments.

22 chromosomes of realistic lengths, simulated under Wright-Fisher model (top) and coalescent (bottom), compared to the analytical expectation under Eqs (1) and (2) in S3 Appendix. Effective population size 10,000, sample size A) 5000, B) 2500, C) 1000, D) 500. Minimum IBD segment length of 1 centimorgan.

S4 Figure Number of IBD segments between pairs of individuals versus total length of shared IBD segments, under the Gutenkunst et. al. (2009) [69] out-of-Africa model. 22 chromosomes of realistic lengths, simulated under Wright-Fisher model (top) and coalescent (bottom), compared to the analytical expectation under Eqs (1) and (2) in S3 Appendix. The African, European, and Asian populations had 1000 haploid samples each.

S5 Figure Computation time of Hudson coalescent, Wright-Fisher, and hybrid models with 100 and 1000 Wright-Fisher generations before switching to the coalescent. Simulations contain from 1 to 22 chromosomes of realistic lengths, using the method described in S1 Appendix, in 500 haploid samples within a diploid population of size 500.

Preface to Chapter 4

In the first manuscript of this thesis, we presented a tool for performing inference at a single locus in a pedigree containing millions of individuals. In the second manuscript we developed a method for performing genome-wide simulations with realistic pairwise relatedness among the simulated samples, even when sample size is large. The natural next step is to combine the strengths of these two methods, and perform genome-wide simulations within large pedigrees. In the following manuscript we see that not only do these simulations capture significant patterns of variation seen within real genetic data, but we show how they may be used directly to guide the design of imputation panels, and outline how they could be used to discover disease associations of rare alleles by directly modelling the fine-scale relatedness within medical cohorts.

Chapter 4

Whole-genome simulations within population-scale pedigrees reflect real patterns of genetic variation

Dominic Nelson[1], Jerome Kelleher[2], Luke Anderson-Trocmé[1], Aaron P. Ragsdale[1], Alexandre Bureau[3], Simon Gravel[1]

In preparation.

1. McGill University and Genome Québec Innovation Centre, McGill University, Montréal, Québec, Canada

2. Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, United Kingdom

3. Département de médecine sociale et préventive, Université Laval, Québec, Québec, Canada

Abstract

With the advent of increasingly high-performance genetic simulation software, large cohorts can now be simulated to aid in the understanding of demographic and evolutionary history. However, as simulated cohorts become larger and more complex, they require more sophisticated models in order to reflect realistic patterns of relatedness and diversity. Here we present a new framework for performing large-scale genetic simulations within a predefined pedigree. Not only does this allow users to directly specify arbitrary relatedness between simulated individuals, but it also allows simulations to be performed using real genealogical data. We compare simulations of individuals within the population-scale pedigree of Quebec, Canada to real genotype data from those individuals, and use principal component analysis to show that simulations capture complex patterns of variation within the real dataset. We outline how these simulations can inform the design of imputation panels, and discuss the use of whole-genome simulations to detect disease associations of rare variants.

Introduction

Coalescent simulators have been extensively used for simulations of large cohorts due to their computational efficiency and well-developed mathematical theory [1]. However, commonly-used coalescent models such as Hudson's [2, 3] exhibit significant distortions of sample relatedness and the distribution of IBD when sample size is large or when simulating long regions [4, 5]. The msprime coalescent simulation software has recently been extended to allow Wright-Fisher simulations [6], which do not share these biases, and allow large whole-genome datasets to be generated. In spite of these improvements the Wright-Fisher model remains a highly idealized representation of real human pedigrees, which are shaped by complex effects such as assortative mating, inbreeding, and isolation-by-distance [7].

To better understand what effects these have on present-day diversity, we further extend msprime to allow simulations to take place within a predefined pedigree. This has several advantages. First, simulations can make use of an increasing number of large genealogical datasets, some of which contain hundreds of thousands to millions of individuals [8, 9, 10], and which provide detailed insights into recent human evolution. Second, pedigrees with desired characteristics can be generated separately in order to isolate the effects of a particular pedigree structure. In either case pedigrees of any size can be used, with simulations continuing under the Wright-Fisher or coalescent models once the founders of the pedigree have been reached.

To demonstrate the versatility of high-performance pedigree simulations, we first explore in section 4 simulations of individuals within the BALSAC genealogy of Quebec, Canada [9], with real data from those

same individuals, and show that simulations capture major components of real patterns of variation. In section 4 we show how simulations in a real pedigree can inform imputation panel design, by evaluating the proportion of present-day genomes imputable given an arbitrary panel composition.

Many other applications are possible and are outlined in the discussion, but detailed investigation is left for future work. One of the most promising is the detection of rare disease associations in family groups within a large genealogy. Using whole-genome simulations of the affected individuals, it is possible to construct a null model of allele sharing that fully captures the potentially complex relatedness patterns among the probands, and makes evaluating the significance of observed associations straightforward, even capturing multiple-testing without resorting to a somewhat arbitrary genome-wide significance threshold. This threshold itself can be investigated, using (neutral) whole-genome simulations of cohorts of various sizes and relatedness patterns, and for alleles at different frequencies.

Results

We first illustrate the benefits of genetic simulations within genealogies. Model and algorithmic details are outlined in the Methods section.

Comparison to a real dataset

To evaluate how well pedigree simulations are able to reproduce patterns of diversity in real datasets, we explore the major axes of variation in both a real and a simulated cohort using principle component analysis (PCA). Real genotype data were taken from the Genizon Biobank [11], of which 2293 individuals have been connected to the BALSAC genealogical database. Since we expect no correlation in actual genotypes between real and simulated data, we instead compare the structure of the embeddings in principle component (PC) space in the following way. Taking the coordinates of all simulated individuals in a single PC, we compute the r^2 correlation of the resulting vector with a vector of PC values from the real dataset. Even though the directions of variation will be different, since they are driven by the genotypes, if the positions of each real and simulated sample along their respective principle components are well-correlated, then we know that simulations have successfully captured the structure of variation in the real data.

We show the correlation between real and simulated PCs in Figure 4.1. The correlation of the first three PCs is high, at approximately 0.8, and PC 6 in the real data is well-correlated with simulated PC 4. Real PCs 4 and 5 do not appear to be captured by the simulations, and we therefore expect they are driven by sources of variation beyond pedigree structure. Real and simulated PCs 1 through 3 show very strong correlation, while simulated PC 4 and real PC 6 seem not to capture much variation in the majority of

individuals in either dataset, as seen in the large cluster of individuals around the origin. These PCs are driven instead by three outliers - siblings from the Saguenay region of Quebec. The replication of this small cluster in simulations is strong evidence for their ability to capture many of the subtleties of real datasets.

UMAP

To more easily visualize patterns of diversity in simulated cohorts, we use PCA combined with the UMAP dimension-reduction method applied to the top 20 PCs. This has been shown to be sensitive to fine-scale population structure, while preserving global structure better than PCA alone [12].

While Wright-Fisher simulations have been shown to be more robust to increasing sample size and segment length, they still lack the fine-scaled structure of a real population. To illustrate the magnitude of the difference, we compared simulated individuals to data from the Genizon biobank, containing genotyped individuals from the province of Quebec, Canada. The results are shown in Figure 4.1 and show strong agreement between real and simulated datasets.

Imputation

When real large pedigrees are available, pedigree simulations allow detailed investigation of imputation quality for arbitrary imputation panel compositions. This can be invaluable when performing a cost/benefit analysis of panel size and sampling scheme.

For imputation to be possible, the individual to be imputed must be IBD with at least one member of the imputation panel at the imputed region, and being IBD with multiple panel members allows more accurate imputation. Since tree sequences contain the ancestry of all individuals across the whole region, simulations within real pedigrees reflect the expected IBD patterns in pedigree individuals. In order to evaluate the imputation power of a given panel, we therefore simply simulate the panel individuals along with a representative set of individuals with which to evaluate IBD sharing with the panel.

Figure 4.2 shows a comparison of the imputation power of two different randomly-sampled panels, containing 1,000 and 10,000 individuals. We evaluate imputation power by examining the percentage of present-day genomes which are IBD with a given number of panel members. In the smaller panel we see that over 40% of present-day genomes are not expected to be imputable, which changes to under 30% in the larger panel. We also see a shift to a more present-day genomes being IBD with a higher number of panel members, leading to more accurate imputation. Similar comparisons can be made for arbitrary panel sampling strategies, and power can be evaluated separately on any collection of present-day individuals. This can give valuable insight when weighing different recruitment strategies for building potentially costly imputation panels.

Performance

Simulating within a fixed pedigree also significantly speeds up simulations of multiple chromosomes. Multiple chromosomes can currently be modelled in `msprime` using recombination hotspots, and whole-genome simulations of tens of thousands of individuals are feasible. However, the computational time is quadratic in the length of the genomic region being simulated, leading to high computational demand [1]. With a fixed pedigree, however, we no longer need to simulate the entire genome as a single contiguous region. Chromosomes are inherited independently conditional on the population pedigree, and so can be simulated separately if the pedigree is fixed. While scaling remains quadratic within individual chromosomes, it becomes linear in number of chromosomes. As shown in Figure 4.3, this leads to large efficiency gains in whole-genome simulations, and allows further gains through parallelization, either on a personal computer or a compute cluster.

Methods

In order to maintain efficiency when simulating within a fixed pedigree, we developed a simulation algorithm based on Wright-Fisher simulations within `msprime`. Pseudocode describing the algorithm is provided in Section 7.5.1 but we outline the steps here. As in Wright-Fisher simulations, we begin with a collection of sample lineages and simulate the ancestral history of these lineages backwards in time. When two lineages share a common ancestor, coalescence occurs within any overlapping genetic regions. Back-and-forth recombination assigns genetic material alternately between grand-parental lineages. Demographic events such as migration and population-size changes are not explicitly modelled, as we assume the pedigree is complete, and so already captures these demographic features.

We require a modified algorithm for two reasons. First, in the Wright-Fisher model generations are discrete and non-overlapping, whereas in real pedigrees this is not the case. Because of this we cannot progress backwards-in-time by stepping through generations, and instead explicitly simulate through all individuals in the pedigree, sorted in reverse chronological order. We do this by maintaining a priority queue of individuals, always taking the most recent, checking for coalescence events within them, then recombining their lineages into each of their parents. The parents are then inserted into the queue, and we proceed to simulate the next-most-recent individual. In this way we guarantee that all individuals are simulated in the proper order, with maintaining the priority queue the only overhead. Since we only ever read out the most-recent individual, we can use an efficient heap structure for this purpose. A detailed overview of the simulation algorithm is given in the supplement.

The second modification we make is to explicitly model diploid individuals. While the Wright-Fisher model of `msprime` implements diploid recombination, lineages are treated independently when parents are drawn. In our pedigree simulations each ancestor is properly modelled as having two fixed lineages. In the future this could be extended to arbitrary ploidy, but doing so would require significant changes to the existing algorithms, and has not yet been implemented.

Discussion and future work

There are several extensions that expand the potential applications of pedigree simulations, and which are well-suited for the pedigree simulation algorithm. A current priority is to model the sex chromosomes. Implementing this, while not trivial, conceptually is straightforward. All that is required is to assign sexes to pedigree individuals, and add checks so that sex chromosomes are inherited appropriately, and for example that recombination in the X chromosomes happens only in females (outside of the pseudo-autosomal region). This extension would allow the investigation of the effects of sex-biased admixture, and generally allow the inclusion of the sex chromosomes in any simulation-based analysis.

Other sexual dimorphisms can be modelled as well. Recombination varies significantly between the sexes, leaving distinct signatures in IBD patterns among close relatives depending on the sex of their shared ancestors [13]. Including sex-specific recombination maps in pedigree simulations would allow detailed IBD analyses at a population scale, and an investigation of the effects of signatures of sex-specific local ancestry.

Different mutational models are also possible. Currently mutations are drawn proportionately to coalescent tree branch lengths, which are measured in generations. However, this also assumes the existence of a 1-1 mapping between number of generations and calendar time. In pedigrees this is in general not possible, as due to inbreeding there may be multiple paths between an individual and one of their ancestors, in particular more distant ones. These paths may not all contain the same number of intermediate ancestors, meaning they do not span the same number of generations, despite the fact that each ancestor lived at a specific calendar time. Being able to distinguish these two measures of ancestral distance is useful because certain types of mutations accumulate according to calendar time, and others according to the number of generations. For example, because in human females their sex cells do not divide throughout their life, the majority of mutations are accumulated at a per-generation rate. In human males however, sex cells divide continuously, and so mutations accumulate over the life of the individual as a function of calendar time.

Some limitations remain. We currently do not model pedigree errors, which are inevitable in large genealogical datasets. Missing data limits our ability to capture subtle patterns of variation, and false paternity and adoptions will also lead to a divergence of real and simulated genetic relatedness. Despite

these challenges, pedigree simulations open up the possibility of directly modelling fine-scale relatedness at large scale, a need which continues to increase as biobank-scale datasets are created and existing ones continue to grow.

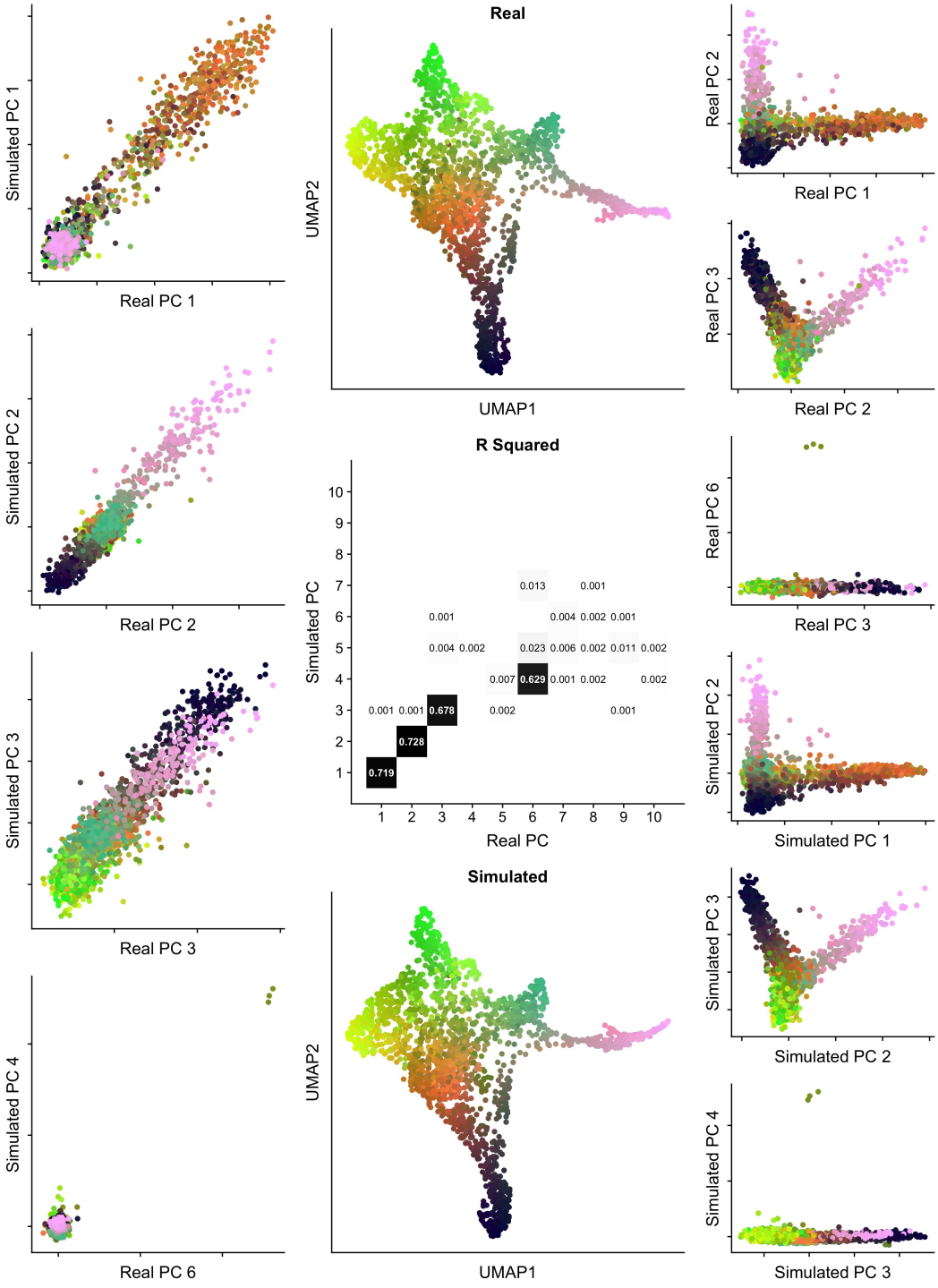


Figure 4.1: Comparison of PCs of real genotypes of 2293 individuals, who have been connected to the BALSAC genealogical database, to PCs of simulated genotypes from the same individuals using our pedigree simulation method. Left column: comparison of most-correlated real and simulated PCs. Middle column, top and bottom: UMAP dimension reduction of the top 20 real and simulated PCs. Middle column, centre: r-squared correlation of real and simulated PCs. Right column: 2D views of real and simulated PC space. Colours are the result of converting 3D UMAP coordinates into RBG colorspace.

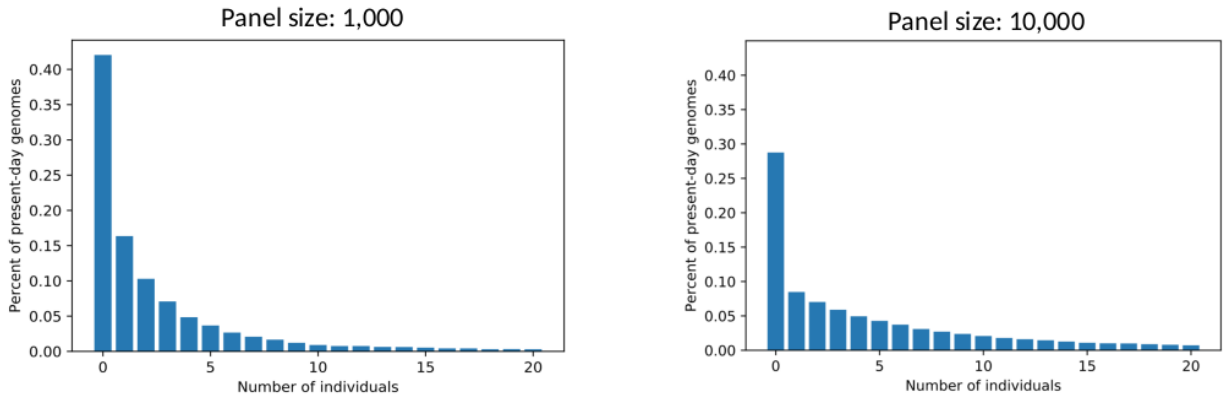


Figure 4.2: Imputation power of two different randomly-sampled panel sizes, shown as the percentage of present-day genomes IBD with a given number of panel individuals. Imputation power was computed by simulating individuals in the panel and counting regions of individual genomes as imputable if their lineages coalesce with those of a panel member within the genealogy.

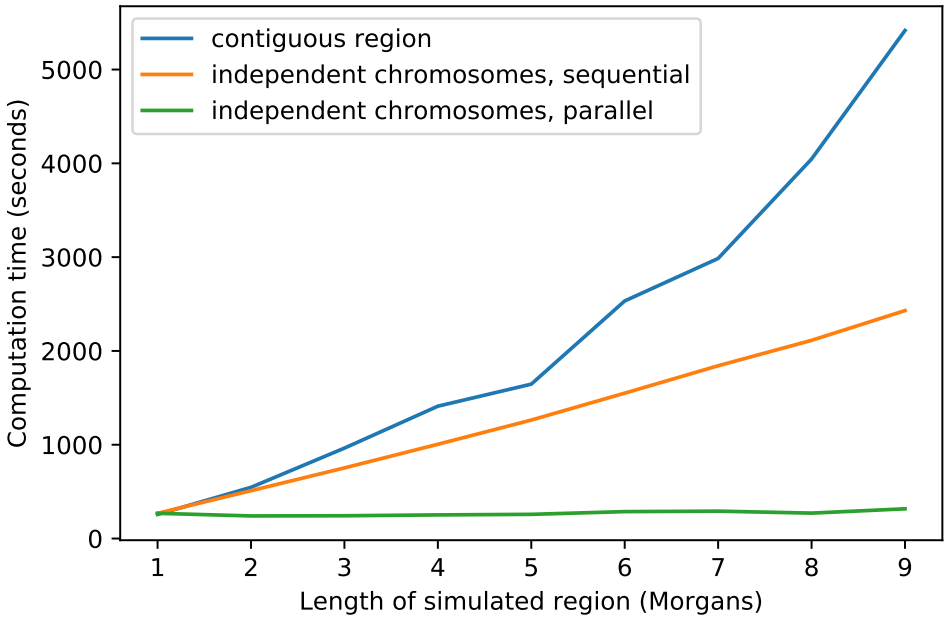


Figure 4.3: Performance of pedigree simulations of multiple chromosomes 1 Morgan in length in 4134 diploid individuals connected to the BALSAC genealogical database. When the top of the pedigree was reached, simulations continued in a single Wright-Fisher population of size 10,000. Compares simulations of a single contiguous region, split into chromosomes using a recombination map; simulations of each chromosome performed independently in sequence; and simulations of each chromosome performed independently in parallel.

Bibliography

[1] Jerome Kelleher, Alison M. Etheridge, and Gilean McVean. "Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes". In: *PLoS Computational Biology* 12.5 (2016), pp. 1–39. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1004842.

[2] Richard R. Hudson. "Properties of a neutral allele model with intragenic recombination". In: *Theoretical Population Biology* 23.2 (1983), pp. 183–201. ISSN: 10960325. DOI: 10.1016/0040-5809(83)90013-8.

[3] Richard R. Hudson. "Gene genealogies and the coalescent process". In: *Futuyma, D. and Antonovics, J. (eds), Oxford Surveys in Evolutionary Biology*. Vol. 7. 1990, pp. 1–44.

[4] Anand Bhaskar, Andrew G Clark, and Yun S Song. "Distortion of genealogical properties when the sample is very large." In: *Proceedings of the National Academy of Sciences of the United States of America* 111.6 (2014), pp. 2385–90. ISSN: 1091-6490. DOI: 10.1073/pnas.1322709111. arXiv: arXiv:1308.0091v1.

[5] John Wakeley et al. "Gene genealogies within a fixed pedigree, and the robustness of kingman's coalescent". In: *Genetics* 190.4 (2012), pp. 1433–1445. ISSN: 00166731. DOI: 10.1534/genetics.111.135574.

[6] Dominic Nelson et al. "Coupling Wright-Fisher and coalescent dynamics for realistic simulation of population-scale datasets". In: *bioRxiv* (2019), p. 674440. DOI: 10.1101/674440.

[7] Gideon S. Bradburd and Peter L. Ralph. "Spatial Population Genetics: It's About Time". In: *Annual Review of Ecology, Evolution, and Systematics* 50.1 (2019). ISSN: 1543-592X. DOI: 10.1146/annurev-ecolsys-110316-022659. arXiv: 1904.09847.

[8] Joanna Kaplanis et al. "Quantitative analysis of population-scale family trees with millions of relatives". In: *Science* 360.April (2018), pp. 171–175.

[9] BALSAC. *BALSAC Population Database: 2016-2017 Annual Report.* http://balsac.uqac.ca/english/files/2018/01/BALSAC_RA2017_EN_page_WEB_v2-1.pdf. Accessed April 8, 2018. 2018.

[10] David O. Arnar and Runolfur Palsson. "Genetics of common complex diseases: a view from Iceland". In: *European Journal of Internal Medicine* 41 (2017), pp. 3–9. ISSN: 18790828. DOI: 10.1016/j.ejim.2017.03.018.

[11] Genome Quebec. *Genizon Biobank*. http://www.genomequebec.com/genizon-biobank/. Accessed January 7, 2020. 2020.

[12] Alex Diaz-Papkovich et al. “UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts”. In: *PLoS Genetics* 15.11 (2019), pp. 1–24. ISSN: 15537404. DOI: 10.1371/journal.pgen.1008432.

[13] Madison Caballero et al. “Surprising impacts of crossover interference and sex-specific genetic maps on identical by descent distributions”. In: *bioRxiv* (2019), p. 527655. DOI: 10.1101/527655.

Chapter 5

General Discussion

Genealogies have long been an integral part of genetic association studies, and continue to produce insights into the relationship between genetics and disease. The development of cheaper whole-genome sequencing and correspondingly larger cohort sizes led to the growth of GWAS as a dominant method of discovering disease associations, which avoided modelling the complexities of detailed relatedness between individuals in favour of more general population-scale demographic models, in order to make analysis of such large datasets feasible. The growth and availability of population-scale pedigrees, along with the development of highly efficient tools for large-scale genealogical inference, now raise the possibility of again modelling the genealogical relationship between cohort individuals, while keeping pace with the ever-growing size of modern medical cohorts.

There are formidable challenges to achieving this goal. Population-scale pedigrees have remained beyond the scale of traditional inference tools [3], and methods for association-testing within large inferred genealogies have not yet been developed to our knowledge. To adapt to modern tools and genomics studies we need efficient pedigree-analysis tools, and methods for validating and extending large-scale genealogical inference.

The work presented here is a significant step towards that goal. In Chapter 1 we discussed ISGen, a software package for estimating the ancestral origin and present-day distribution of rare alleles within large pedigrees. Using importance sampling to improve the efficiency of Monte Carlo simulations, ISGen scales well beyond the capabilities of MCMC-based inference tools, allowing allele frequency estimates in pedigrees containing millions of individuals, something which to our knowledge has never before been possible.

However, despite the effort spent optimizing ISGen, part of its efficiency is due to a careful and somewhat limited scope. Inference is only possible within a single shared haplotype, and only for alleles rare enough to have likely been introduced to the population through a single pedigree founder. While large-scale datasets will uncover many more rare alleles, there is clearly utility to be gained from integrating data across whole chromosomes or genomes.

Luckily, tools for performing ancestral inference at whole-genome scale are already available, and continue to be actively developed. As discussed in the Introduction, ARGweaver [54], Relate [56], and `tsinfer` [59] are capable of inferring whole-genome gene genealogies, numbering in the hundreds of thousands in the case of `tsinfer` [59].

However, there remain challenges in the development of such tools, one of which being the systematic evaluation of their accuracy and potential biases. An invaluable method for evaluating such tools is to test them on simulated data, where their output can be compared against the known simulation parameters and population history. While this has been straightforward in the past, when simulated datasets were smaller, large cohorts present new challenges to simulation tools, requiring both high computational efficiency and significantly more sophisticated models to capture the subtle relatedness patterns inevitably present in large groups of individuals [9].

The analysis performed in Chapter 2 highlights the limitations of state-of-the-art coalescent simulators when simulating at whole-genome scale. The assumptions built into the coalescent model, while valid at smaller scales, lead to large biases in relatedness patterns as sample sizes and simulated sequence length continue to grow.

Towards the goal of supporting population-scale genealogical inference, the Wright-Fisher extension to `msprime`, also presented in Chapter 2, provides a new framework for performing realistic simulations of large cohorts, even out-performing the fastest coalescent simulations at whole-genome scale while generating more accurate relatedness between individuals in large simulated cohorts. It generates and stores simulated genomes in the tree sequence format of `msprime` and `tsinfer`, making stored output space-efficient and simplifying the comparison of simulated and inferred genealogies.

As relatedness within cohorts becomes more important, the natural progression beyond a stable, but idealized, Wright-Fisher pedigree, is to simulate within a real pedigree. Several pedigrees totalling over a million individuals each are now available, and greatly expand the possible applications of simulation-based methods. A few such application were described in Chapter 3, such as rare-variant association testing and the evaluation of the imputation power of different cohort designs. We expect many more applications to come with the possibility of performing genome-wide simulations of hundreds of thousands of individuals, with relatedness matching that of a real population and the detailed demographic history which shaped it.

There are of course still limitations to the tools presented in this thesis, and many opportunities for future improvements. First, we have focused predominantly on simulations under a neutral model, and discounted gene conversion, indels, inversion, and more complex chromosomal rearrangements. Models of neutral SNPs could also be improved by incorporating sex-specific mutation models, where males continue to accumulate new mutations over the course of their life. Pedigree simulations also currently depend on the accuracy of the pedigree data, which may not completely reflect genetic relatedness due to adoption or false paternity.

More generally we can compare these tools to forwards-time alternatives, such as the SLiM simulation software [70]. Working forwards-in-time simplifies simulation of selection, but also presents new challenges. Since it is not known ahead of time which ancestral individuals contributed genetic material to the present

day, they must all must be simulated across the whole genetic region of interest. Any which do not ultimately contribute represent wasted computational effort. Simulating backwards-in-time ensures that only relevant ancestral genetic material will be generated.

Other questions require deeper consideration. For example, modelling selection implies that the allele frequency of selected loci over time has an impact on pedigree structure, since fitness measures an individual's reproductive success. When the pedigree is unknown, it can be generated in parallel with simulated ancestral genomes to match the expected distribution of numbers of offspring. How to include selection in simulations when the pedigree is known is less clear. Inheritance of selected alleles now depends on the pedigree structure itself, where alleles with positive fitness effects are more likely to have been inherited from individuals with larger numbers of offspring. Modelling this correlation for multiple selected loci while maintaining linkage patterns across long regions of neutral variation will be challenging, in particular under limits to computational complexity.

Despite these challenges, pedigree simulations are an invaluable tool for further refinement of genealogical inference tools such as ARGweaver, Relate, and `tsinfer`. Beyond providing more realistic simulated cohorts for testing and validation, genealogies inferred from pedigree-based simulations can then be aligned back to the original pedigree itself. This will require substantial research effort, but can ultimately lead to large-scale imputation of ancestral genotypes, similar to the single-locus imputation done by ISGen in Chapter 1, but genome-wide, and using the ascending genealogies of potentially hundreds of thousands of individuals.

Chapter 6

Conclusions and Future Directions

This thesis has presented three new tools, freely available and open-source, which can help genetics researchers to incorporate and study fine-scale genealogical models. Many promising avenues for future work remain, in particular those which explore further applications of the pedigree simulations described in Chapter 4.

An efficient method for whole-genome pedigree simulations opens up several interesting applications beyond aiding genealogical inference efforts. One is to investigate the robustness of the ubiquitous genome-wide significance threshold. While the simplicity of a widely-accepted threshold for genome-wide significance has its advantages, it is also not truly universal. The true number of effective tests can vary with cohort size and composition, as well as with the rarity of the causal allele. As cohorts continue to grow, it is worthwhile to validate the accepted genome-wide significance threshold to fit this new paradigm, in order to fully take advantage of these comprehensive (and costly) sequencing efforts.

The relationship between effective population size and census population size can also be investigated using pedigree simulations. Effective population size can be difficult to interpret in a real population, but the existence of population-scale pedigrees allows some direct comparisons to be made. One possibility is to simply simulate individuals with a pedigree, infer the effective population size from the simulated data, and compare this to the known census population size.

Beyond using real pedigrees, pedigrees constructed with varying values for parameters such as inbreeding, outbreeding, spatial dispersion, or assortative mating could be used to simulate genetic data to determine the effects of these parameters on the inferred effective population size. Another intriguing possibility is to simulate within a real pedigree, infer the effective population size, and then simulate a randomly-mating population of the inferred size. This would allow a comparison of the two simulated datasets to determine possible variation between populations with the same effective population size, and how these variations are informative about pedigree structure and large-scale demographic history.

Another possibility is to use pedigree simulations to greatly simplify significance-testing of disease associations of rare alleles. The genome-wide significance threshold is a function of the correlation structure of variation across the genome, giving the largest p-value likely to be seen purely by chance, as opposed to resulting from a real association between variant and phenotype. With genome-wide pedigree simulations, rather than building a null model for a single locus and correcting it genome-wide, we can build a

genome-wide null model directly.

As an example, take a set of affected individuals who can be placed in a large pedigree. The standard approach to identifying associated variants involves performing a large number of individual tests of association, and then correcting the resulting p-values according to the number of tests performed. If instead we were able to accurately simulate the genomes of the affected individuals under a neutral model, we could simply calculate our association statistic on the simulated data, and using a set of simulations we can get a distribution of the statistic. For example, we could look at how often the affected individuals share an allele. If we simulate 1000 times and allele sharing matches or exceeds observed sharing only 10 times, we can conclude that an association exists at the 99% confidence level. This method is flexible in the choice of statistic, and by performing genome-wide tests avoids the complexities of multiple-testing corrections. This GWAS strategy also accounts for arbitrary relatedness among affected individuals, since it uses a known pedigree when performing simulations.

Biobanks and medical cohorts continue increase in size, and have begun to outgrow traditional methods of analysis. Fine-scale relatedness can no longer be ignored, and new methods are needed with this understanding built-in. The tools and strategies described here aim to provide a practical foundation for continued growth in the field, and directly support the investigation of rare-disease associations, optimal construction of imputation panels, and validation of methods for large-scale genealogical inference.

Bibliography

[1] F. Monaghan and A. Corcos. "On the origins of the Mendelian laws". In: *Journal of Heredity* 75.1 (1984), pp. 67–69. ISSN: 00221503. DOI: 10.1093/oxfordjournals.jhered.a109868.

[2] Alan Medlar et al. "SwiftLink: Parallel MCMC linkage analysis using multicore CPU and GPU". In: *Bioinformatics* 29.4 (2013), pp. 413–419. ISSN: 13674803. DOI: 10.1093/bioinformatics/bts704.

[3] Dominic Nelson et al. "Inferring Transmission Histories of Rare Alleles in Population-Scale Genealogies". In: *American Journal of Human Genetics* 103.6 (2018), pp. 893–906. ISSN: 15376605. DOI: 10.1016/j.ajhg.2018.10.017.

[4] Lucia A. Hindorff et al. "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits". In: *Proceedings of the National Academy of Sciences of the United States of America* 106.23 (2009), pp. 9362–9367. ISSN: 00278424. DOI: 10.1073/pnas.0903103106.

[5] Jacqueline MacArthur et al. "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)". In: *Nucleic Acids Research* 45.D1 (2017), pp. D896–D901. ISSN: 13624962. DOI: 10.1093/nar/gkw1133.

[6] Itsik Pe'er et al. "Estimation of the multiple testing burden for genomewide association studies of nearly all common variants". In: *Genetic Epidemiology* 32.4 (2008), pp. 381–385. ISSN: 07410395. DOI: 10.1002/gepi.20303.

[7] João Fadista et al. "The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants". In: *European Journal of Human Genetics* 24.8 (2016), pp. 1202–1205. ISSN: 14765438. DOI: 10.1038/ejhg.2015.269.

[8] Xiaoyi Gao, Joshua Starmer, and Eden R. Martin. "A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms". In: *Genetic Epidemiology* 32.4 (2008), pp. 361–369. ISSN: 07410395. DOI: 10.1002/gepi.20310.

[9] Vladimir Shchur and Rasmus Nielsen. "On the number of siblings and p-th cousins in a large population sample". In: *Journal of Mathematical Biology* 77.5 (2018), pp. 1–20. ISSN: 14321416. DOI: 10.1007/s00285-018-1252-8.

[10] Brenna M Henn et al. "Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples". In: *PLoS ONE* 7.4 (2012). ISSN: 19326203. DOI: 10.1371/journal.pone.0034267.

[11] Georgios Athanasiadis et al. "Nationwide genomic study in Denmark reveals remarkable population homogeneity". In: *Genetics* 204.2 (2016), pp. 711–722. ISSN: 19432631. DOI: 10.1534/genetics.116.189241.

[12] Po Ru Loh et al. "Efficient Bayesian mixed-model analysis increases association power in large cohorts". In: *Nature Genetics* 47.3 (2015), pp. 284–290. ISSN: 15461718. DOI: 10.1038/ng.3190.

[13] John Wakeley, Léandra King, and Peter R. Wilton. "Effects of the population pedigree on genetic signatures of historical demographic events". In: *Proceedings of the National Academy of Sciences* 113.29 (2016), pp. 7994–8001. ISSN: 0027-8424. DOI: 10.1073/pnas.1601080113.

[14] Sewall Wright. "Evolution in Mendelian populations". In: *Genetics* 16.2 (1931), p. 97.

[15] RA Fisher. *The genetical theory of natural selection.* Clarendon Press, 1930.

[16] Warren J. Ewens. *Mathematical Population Genetics I: Theoretical introduction.* Springer New York, 2004.

[17] James F. Crow and Motoo Kimura. *An Introduction to Population Genetics Theory.* Harper and Row, 1970.

[18] Daniel L. Hartl and Andrew G. Clark. *Principles of Population Genetics.* Sinauer Associates, 1997.

[19] Freddy B. Christiansen. *Theories of Population Variation in Genes and Genomes.* Princeton University Press, 2008.

[20] J F C Kingman. "The coalescent". In: 13 (1982), pp. 235–248.

[21] Richard R. Hudson. "Properties of a neutral allele model with intragenic recombination". In: *Theoretical Population Biology* 23.2 (1983), pp. 183–201. ISSN: 10960325. DOI: 10.1016/0040-5809(83)90013-8.

[22] Fumio Tajima. "Evolutionary relationship of DNA sequenes in finite populations". In: *Genetics* 105 (1983), pp. 437–460.

[23] Peter R. Wilton et al. "Population structure and coalescence in pedigrees: Comparisons to the structured coalescent and a framework for inference". In: *Theoretical Population Biology* 115 (2017), pp. 1–12. ISSN: 10960325. DOI: 10.1016/j.tpb.2017.01.004.

[24] John Wakeley et al. "Gene genealogies within a fixed pedigree, and the robustness of kingman's coalescent". In: *Genetics* 190.4 (2012), pp. 1433–1445. ISSN: 00166731. DOI: 10.1534/genetics.111.135574.

[25] David O. Arnar and Runolfur Palsson. "Genetics of common complex diseases: a view from Iceland". In: *European Journal of Internal Medicine* 41 (2017), pp. 3–9. ISSN: 18790828. DOI: 10.1016/j.ejim.2017.03.018.

[26] Agnar Helgason et al. "An Association Between the Kinship and Fertility of Human Couples". In: *Science* 319.5864 (2008), pp. 813–816. ISSN: 0036-8075. DOI: 10.1126/science.1150232.

[27] Anna Helgadottir et al. "Variants with large effects on blood lipids and the role of cholesterol and triglycerides in coronary disease". In: *Nature Publishing Group* 48.6 (2016), pp. 634–639. ISSN: 1061-4036. DOI: 10.1038/ng.3561.

[28] Struan F.A. Grant et al. "Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes". In: *Nature Genetics* 38.3 (2006), pp. 320–323. ISSN: 10614036. DOI: 10.1038/ng1732.

[29] Hreinn Stefansson et al. "A common inversion under selection in Europeans". In: *Nature Genetics* 37.2 (2005), pp. 129–137. ISSN: 10614036. DOI: 10.1038/ng1508.

[30] M. Skolnick. "The Utah Genealogical Database: a resource for genetic epidemiology." In: *Cancer incidence in defined populations.* Ed. by M Skolnick J Cairns J Lyon. Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 1980, pp. 285–296.

[31] Lee L Bean, Dean L May, and Mark Skolnick. "The Mormon Historical Demography Project". In: *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 11.1 (1978), pp. 45–53. DOI: 10.1080/01615440.1978.9955216.

[32] Shane Lloyd et al. "Mental health disorders are more common in colorectal cancer survivors and associated with decreased overall survival". In: *American Journal of Clinical Oncology: Cancer Clinical Trials* 42.4 (2019), pp. 355–362. ISSN: 1537453X. DOI: 10.1097/COC.0000000000000529.

[33] Lisa A. Cannon-Albright et al. "Relative risk for Alzheimer disease based on complete family history". In: *Neurology* 92.15 (2019), e1745–e1753. ISSN: 1526632X. DOI: 10.1212/WNL.0000000000007231.

[34] Makenzie L. Hawkins et al. "Endocrine and Metabolic Diseases Among Colorectal Cancer Survivors in a Population-Based Cohort". In: *Journal of the National Cancer Institute* 112.1 (2020), pp. 78–86. ISSN: 14602105. DOI: 10.1093/jnci/djz040.

[35] Sarah Hummel et al. "The contribution of the rs55705857 G allele to familial cancer risk as estimated in the Utah population database". In: *BMC Cancer* 19.1 (2019), pp. 1–6. ISSN: 14712407. DOI: 10.1186/s12885-019-5381-2.

[36] Niels van den Berg et al. "Longevity defined as top 10% survivors and beyond is transmitted as a quantitative genetic trait". In: *Nature Communications* 10.1 (2019). ISSN: 20411723. DOI: 10.1038/s41467-018-07925-0.

[37] BALSAC. *BALSAC Population Database: 2016-2017 Annual Report.* http://balsac.uqac.ca/english/files/2018/01/BALSAC_RA2017_EN_page_WEB_v2-1.pdf. Accessed April 8, 2018. 2018.

[38] Evelyne Heyer et al. “Estimating Y Chromosome Specific Microsatellite Mutation Frequencies using Deep Rooting Pedigrees”. English. In: *Human Molecular Genetics* 6.5 (1997), pp. 799–803. ISSN: 0964-6906.

[39] E Heyer et al. “Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees.” English. In: *American journal of human genetics* 69.5 (2001), pp. 1113–1126. ISSN: 0002-9297.

[40] Stephan Peischl et al. “Relaxed selection during a recent human expansion”. In: *Genetics* 208.2 (2018), pp. 763–777. ISSN: 19432631. DOI: 10.1534/genetics.117.300551.

[41] Roy-Gagnon MH et al. “Genomic and genealogical investigation of the French Canadian founder population structure.” English. In: *Human genetics* 129.5 (2011), pp. 521–531. ISSN: 0340-6717.

[42] Claudia Moreau et al. “Native American Admixture in the Quebec Founder Population”. In: *PLoS ONE* 8.6 (2013), pp. 1–9. ISSN: 19326203. DOI: 10.1371/journal.pone.0065507.

[43] C. Moreau et al. “Deep Human Genealogies Reveal a Selective Advantage to Be on an Expanding Wave Front”. In: *Science* 334.6059 (2011), pp. 1148–1150. ISSN: 0036-8075. DOI: 10.1126/science.1212880.

[44] Héloïse Gauvin et al. “Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population”. No Linguistic Content. In: *Eur J Hum Genet European Journal of Human Genetics* 22.6 (2014), pp. 814–821. ISSN: 1018-4813.

[45] Philippe Chetaille et al. “Mutations in SGOL1 cause a novel cohesinopathy affecting heart and gut rhythm”. In: *Nature Genetics* 46.11 (2014), pp. 1245–1248. ISSN: 15461718. DOI: 10.1038/ng.3113.

[46] Emmanuel Milot et al. “Mother’s curse neutralizes natural selection against a human genetic disease over three centuries”. In: *Nature Ecology and Evolution* 1.9 (2017), pp. 1400–1406. ISSN: 2397334X. DOI: 10.1038/s41559-017-0276-6.

[47] Genome Quebec. *Genizon Biobank*. http://www.genomequebec.com/genizon-biobank/. Accessed January 7, 2020. 2020.

[48] Philip Awadalla et al. “Cohort profile of the CARTaGENE study: Quebec’s population-based biobank for public health and personalized genomics”. In: *International Journal of Epidemiology* 42.5 (2013), pp. 1285–1299. ISSN: 03005771. DOI: 10.1093/ije/dys160.

[49] Joanna Kaplanis et al. “Quantitative analysis of population-scale family trees with millions of relatives”. In: *Science* 360.April (2018), pp. 171–175.

[50] Penelope Maza. “Adoption Trends: 1944-1975”. In: *Child welfare research notes* 9 (1984), pp. 1–11.

[51] Kermyt G. Anderson. "How Well Does Paternity Confidence Match Actual Paternity? Evidence from Worldwide Nonpaternity Rates". In: *Current Anthropology* 47.3 (2006), pp. 513–520. ISSN: 1098-6596. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

[52] Turi E. King and Mark A. Jobling. "Founders, drift, and infidelity: The relationship between y chromosome diversity and patrilineal surnames". In: *Molecular Biology and Evolution* 26.5 (2009), pp. 1093–1102. ISSN: 07374038. DOI: 10.1093/molbev/msp022.

[53] Eunjung Han et al. "Clustering of 770,000 genomes reveals post-colonial population structure of North America". In: *Nature Communications* 8 (2017). ISSN: 20411723. DOI: 10.1038/ncomms14238.

[54] Matthew D. Rasmussen et al. "Genome-Wide Inference of Ancestral Recombination Graphs". In: *PLoS Genetics* 10.5 (2014). ISSN: 15537404. DOI: 10.1371/journal.pgen.1004342. arXiv: 1306.5110 [q-bio.PE].

[55] RC Griffiths and P Marjoram. "Progress in population genetics and human evolution". In: (1997).

[56] Leo Speidel et al. "A method for genome-wide genealogy estimation for thousands of samples". In: *Nature Genetics* 51.9 (2019), pp. 1321–1329. ISSN: 15461718. DOI: 10.1038/s41588-019-0484-x.

[57] Na Li and Matthew Stephens. "Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data". In: *Genetics* 165.4 (2003), pp. 2213–2233. ISSN: 00166731. DOI: 10.1534/genetics.104.030692.

[58] Jerome Kelleher, Alison M. Etheridge, and Gilean McVean. "Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes". In: *PLoS Computational Biology* 12.5 (2016), pp. 1–39. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1004842.

[59] Jerome Kelleher et al. "Inferring whole-genome histories in large population datasets". In: *Nature Genetics* 51.9 (2019), pp. 1330–1338. ISSN: 1061-4036. DOI: 10.1038/s41588-019-0483-y.

[60] Nancy Chen et al. "Allele frequency dynamics in a pedigreed natural population". In: *Proceedings of the National Academy of Sciences of the United States of America* 116.6 (2019), pp. 2158–2164. ISSN: 10916490. DOI: 10.1073/pnas.1813852116.

[61] Richard R Hudson. "Generating samples under a Wright-Fisher neutral model of genetic variation". In: *Bioinformatics* 18.2 (2002), pp. 337–338. ISSN: 13674803. DOI: 10.1093/bioinformatics/18.2.337.

[62] Benjamin F Voight et al. "A map of recent positive selection in the human genome." In: *PLoS biology* 4.3 (2006), e72. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0040072.

[63] Christopher S Carlson et al. “Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium”. In: *The American Journal of Human Genetics* 74.1 (2004), pp. 106–120. ISSN: 00029297. DOI: 10.1086/381000.

[64] Heng Li and Richard Durbin. “Inference of human population history from individual whole-genome sequences”. In: *Nature* 475.7357 (2011), pp. 493–496. ISSN: 0028-0836. DOI: 10.1038/nature10231.

[65] G. A. T. McVean and N. J. Cardin. “Approximating the coalescent with recombination”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1459 (2005), pp. 1387–1393. ISSN: 0962-8436. DOI: 10.1098/rstb.2005.1673.

[66] Gary K. Chen, Paul Marjoram, and Jeffrey D. Wall. “Fast and flexible simulation of DNA sequence data”. In: *Genome Research* 19.1 (2009), pp. 136–142. ISSN: 10889051. DOI: 10.1101/gr.083634.108.

[67] Anand Bhaskar, Andrew G Clark, and Yun S Song. “Distortion of genealogical properties when the sample is very large.” In: *Proceedings of the National Academy of Sciences of the United States of America* 111.6 (2014), pp. 2385–90. ISSN: 1091-6490. DOI: 10.1073/pnas.1322709111. arXiv: arXiv:1308.0091v1.

[68] Joanna L. Davies et al. “On recombination-induced multiple and simultaneous coalescent events”. In: *Genetics* 177.4 (2007), pp. 2151–2160. ISSN: 00166731. DOI: 10.1534/genetics.107.071126.

[69] Ryan N. Gutenkunst et al. “Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data”. In: *PLoS Genetics* 5.10 (2009). ISSN: 15537390. DOI: 10.1371/journal.pgen.1000695.

[70] Benjamin C. Haller and Philipp W. Messer. “SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model”. In: *Molecular Biology and Evolution* 36.3 (2019), pp. 632–637. ISSN: 15371719. DOI: 10.1093/molbev/msy228.

[71] Philippe Chetaille et al. “Mutations in SGOL1 cause a novel cohesinopathy affecting heart and gut rhythm”. In: *Nature Genetics* 46.11 (2014), pp. 1245–1248. ISSN: 15461718. DOI: 10.1038/ng.3113.

[72] Héloïse Gauvin et al. “Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population”. In: *European Journal of Human Genetics* 22.6 (2014), pp. 814–821. ISSN: 1018-4813.

[73] Claudia Moreau et al. “Native American Admixture in the Quebec Founder Population”. In: *PLoS ONE* 8.6 (2013), e65507.

[74] M-H Roy-Gagnon et al. “Genomic and genealogical investigation of the French Canadian founder population structure.” In: *Human genetics* 129.5 (2011), pp. 521–531. ISSN: 0340-6717.

[75] Quebec Reference Sample. (2010). *Quebec reference sample: Population genetics and genetic epidemiology in Quebec*. http://www.quebecgenpop.ca/. Accessed May 14, 2018. 2010.

[76] Philip Awadalla et al. "Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics". In: *International Journal of Epidemiology* 42.5 (2013), pp. 1285–1299. ISSN: 03005771. DOI: 10.1093/ije/dys160.

[77] BL Browning and SR Browning. "Improving the accuracy and efficiency of identity-by-descent detection in population data." In: *Genetics* 194.2 (2013), pp. 459–471. ISSN: 0016-6731.

[78] Rajan Srinivasan. *Importance Sampling: Applications in Communications and Detection*. Springer-Verlag, 2002.

[79] Léandra King, John Wakeley, and Shai Carmi. "A non-zero variance of Tajima's estimator for two sequences even for infinitely many unlinked loci". In: *Theoretical Population Biology* 122 (2018), pp. 22–29. ISSN: 10960325. DOI: 10.1016/j.tpb.2017.03.002.

[80] W. G. Hill and Alan Robertson. "Linkage disequilibrium in finite populations". In: *Theoretical and Applied Genetics* 38.6 (1968), pp. 226–231. ISSN: 00405752. DOI: 10.1007/BF01245622.

[81] Aaron P Ragsdale and Simon Gravel. "Unbiased Estimation of Linkage Disequilibrium from Unphased Data". In: *Molecular Biology and Evolution* (Nov. 2019). ISSN: 0737-4038.

[82] Alun Thomas, Mark H. Skolnick, and Cathryn M. Lewis. "Genomic mismatch scanning in pedigrees". In: *Mathematical Medicine and Biology* 11.1 (1994), pp. 1–16. ISSN: 14778599. DOI: 10.1093/imammb/11.1.1.

[83] Chad D. Huff et al. "Maximum-likelihood estimation of recent shared ancestry (ERSA)". In: *Genome Research* 21.5 (2011), pp. 768–774. ISSN: 10889051. DOI: 10.1101/gr.115972.110.

[84] Yaniv Erlich et al. "Identity inference of genomic data using long-range familial searches". In: *Science (New York, N.Y.)* 362.6415 (2018), pp. 690–694. ISSN: 10959203. DOI: 10.1126/science.aau4832.

[85] Kevin P. Donnelly. "The probability that related individuals share some section of genome identical by descent". In: *Theoretical Population Biology* 23.1 (1983), pp. 34–63. ISSN: 10960325. DOI: 10.1016/0040-5809(83)90004-7.

[86] Pier Francesco Palamara. "ARGON: fast, whole-genome simulation of the discrete time Wright-fisher process". In: *Bioinformatics* 32.19 (June 2016), pp. 3032–3034. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw355.

Chapter 7

Appendices and Supplemental Material

7.1 Chapter 2 Appendices

7.1.1 Symbol Glossary

ω Minor allele frequency in ancestral source population

$N_{founders}$ Number of founders in the genealogy

a Ancestral (founder) origin of minor allele

$\mathcal{A}$ Set of all founders in the genealogy

c The set of haplotypes within genealogically-connected individuals which have been observed to be minor

S The (observed) event that haplotypes c carry the minor allele

Γ A simulated inheritance path ascending from the minor alleles within c

A A random variable representing the founder who carried the minor allele

$\mathbb{1}_a(\Gamma)$ Indicator function denoting if Γ coalesces to ancestor a

M Number of Monte Carlo iterations

p Original (unbiased) probability distribution of inheritance paths

q Importance sampling (biased) probability distribution of inheritance paths

α Number of allele transmissions in path Γ

β Number of allele transmissions in path Γ with only one valid maternal/paternal path consistent with coalescence

γ Number of times a homozygote inconsistent with coalescence could have been created during the climbing process

F Random variable representing the minor allele frequency in the population, independent of genealogical information

f Observed value of the minor allele frequency in the population

$\partial\Gamma$ Boundary of Γ (first-generation descendants who do not carry a minor allele)

ϕ_k Binomial success probability of ancestors in probability bin i being the true generating ancestors

τ_k Total number of ancestors in bin i

x_k Number of true generating ancestors in bin i

E_Γ Expectation summed over inheritance paths Γ_j

$B_i \sim b_i(t_i)$ The contribution of individual i to global minor allele frequency given they have a single parent simulated to carry t_i alleles

$Y_i \sim y_i(t_i)$ The contribution of individual i to global minor allele frequency given they carry t_i alleles themselves

$\delta_{i,j}$ Kronecker delta function

K True number of carriers in population

N True number of individuals in the population

n Size of sample taken from population (of size N)

k Number of observed carriers in sample n

$H(k; N, n, K)$ Hypergeometric distribution

$\nu_{i,self}$ Number of alleles carried by individual i

$\nu_{i,parent}$ Number of alleles carried by the parent (who is simulated to have carried an allele) of individual i

Λ Event that all minor allele lineages coalesce in the genealogy

$\mathbb{1}_\Lambda(\Gamma)$ Indicator function denoting whether Γ coalesces to a single ancestor

R_m Random variable representing minor allele frequency in an arbitrary region m

r_m Realized value of R_m

$\hat{r}_{m,\mathbf{kin}}$ Kinship-based estimate of regional allele frequency r_m

$\hat{r}_{m,\textbf{kin, corrected}}$ Kinship-based estimate of regional allele frequency r_m, corrected to be conditional on global frequency of minor allele.

h Number of meioses since the most recent common ancestor (MRCA) of the carriers

L Length in Morgans of longest haplotype shared among all carriers of the minor allele

l Observed value of L

7.1.2 Jointly Modelling Individuals Inside and Outside of the Genealogy

We explained in the main text how to compute the posterior probability $P(a|S)$ of ancestor a being the ancestral carrier given the observed event S that the observed carriers received the minor alleles. We want to use the refined posterior $P(a|S, F = f)$, where F is the random variable denoting the minor allele frequency in individuals not linked to the genealogy. As before, this will be computed from the likelihood using Bayes theorem and a flat prior on all ancestors $P(a) = \frac{1}{|\mathcal{A}|}$. Letting $\mathcal{A}$ represent the set of all founding individuals.

$$P(a|S, F = f) = \frac{P(S, F = f|a)P(a)}{\sum_{a' \in \mathcal{A}} P(S, F = f|a')P(a')} \tag{7.1}$$

$$= \frac{P(S, F = f|a)}{\sum_{a' \in \mathcal{A}} P(S, F = f|a')}. \tag{7.2}$$

Now recall that $\mathbb{1}_a(\Gamma)$ indicates whether a simulated inheritance path Γ coalesces to founding ancestor a, so that $P(S|\Gamma, a) = \mathbb{1}_a(\Gamma)$, and the probability $P(\Gamma)$ of an inheritance path is independent of a, that is, $P(\Gamma|a) = P(\Gamma)$. We then have

$$\begin{aligned} P(S, F = f|a) &= \sum_{\Gamma} P(S, F = f|\Gamma, a)P(\Gamma|a) \\ &= \sum_{\Gamma} P(F = f|\Gamma, S, a)P(S|\Gamma, a)P(\Gamma|a) \\ &= \sum_{\Gamma} P(F = f|\Gamma, S, a)\mathbb{1}_a(\Gamma)P(\Gamma). \end{aligned} \tag{7.3}$$

Under the importance sampling scheme described in the main text, we can rewrite this estimate as

$$\begin{aligned} P(S, F = f|a) &= E_{\Gamma}\big[P(F = f|\Gamma, S, a)\mathbb{1}_a(\Gamma)\big] \\ &\simeq \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}_a(\Gamma_j)\,\frac{p(\Gamma_j)}{q(\Gamma_j)}P(F = f|\Gamma_j, S, a). \end{aligned} \tag{7.4}$$

This expression can then be substituted into (7.2) to provide an importance sampling estimate of $P(a|S, F = f)$.

7.1.3 Efficiently Estimating the Probability of the Observed Allele Frequency

In the main text and Fig. 2.3, we argued that the probability distribution of the population allele frequency $P(F|\Gamma)$ can be estimated by performing a sum over the contributions of individuals in the path boundary $\partial\Gamma$, if individuals within Γ are all carry the minor allele.

Because the alleles of individuals in $\partial\Gamma$ are left unassigned during the climbing process that generated Γ, their contributions to the number of minor alleles in the population first depends on whether or not they received minor alleles from individuals in Γ. For simplicity of exposition we assume that each boundary individual has only one parent in the tree, although similar derivations can be made when both parents are in Γ. Since this is a rare occurrence, ISGen currently treats each individual in the boundary of the tree as if it had a single parent in Γ.

For each individual i in $\partial\Gamma$, we first denote by $\nu_{i,parent}$ the number of copies of the minor allele their parent in Γ was simulated to have carried, and by $\nu_{i,self}$ the number of copies of the minor allele they may carry themselves. Let Y_i be the number of copies of the minor allele that i contributes to the present-day population, and $y_i[\nu_{i,self}]$ the distribution of Y_i given that i carried $\nu_{i,self}$ copies of the minor alleles:

$$Y_i|\nu_{i,self} \sim y_i[\nu_{i,self}].$$

We estimate this distribution using a single set of genealogy-wide allele-dropping simulations.

Then, assuming that $i \in \partial\Gamma$, let B_i denote the number of *minor* alleles that i contributes to the present-day population. Given the single-founder assumption, the minor allele frequency in a population of size N (excluding alleles inherited through Γ) is

$$F = \frac{1}{N} \sum_{i \in \partial\Gamma} B_i. \tag{7.5}$$

We estimate the expected B_i by conditioning on the possible transmissions. Let $b_i[\nu_{i,parent}]$ be the conditional distribution of B_i given that the parent of i in Γ carries $\nu_{i,parent}$ alleles:

$$B_i|\nu_{i,parent} \sim b_i[\nu_{i,parent}].$$

If we neglect the probability of inheriting a minor allele from the parent outside Γ, the conditional distribu-

tions $b_i[\nu_{i,parent}]$ and $y_i[\nu_{i,self}]$ follow

$$\begin{aligned} b_i[0](B_i) &\simeq \delta_{0,B_i} \\ b_i[1](B_i) &\simeq \frac{1}{2}\delta_{0,B_i} + \frac{1}{2}y_i[1](B_i) \\ b_i[2](B_i) &\simeq y_i[1](B_i). \end{aligned}$$

The distribution of F can be then calculated using Eq. (7.5) via the convolution of the corresponding $b_i[\nu_{i,parent}]$. In this way, once we have simulated y_i for all individuals i in the genealogy, we can quickly estimate the distribution of F for any Γ encountered in our Monte Carlo simulations, giving a huge gain in efficiency over a large number of simulated inheritance paths. A comparison of this method to allele-dropping simulations is shown in Fig. 7.3.

Finite Sample Estimates of the Allele Frequency

In practice, the population allele frequency in individuals not connected to the genealogy is estimated from a sample of the population. We first denote the population size by N, and let the total number of minor alleles (observed and unobserved) in the population be represented by K.

In the main text, a trajectory Γ only contributes to the likelihood if it coalesces to the contributing founder, an event we label as Λ in this section to simplify notation. Given Λ, the likelihood of an inheritance path Γ giving rise to the observed number of carriers $k = fN$ in a population sample of size n is given by summing over all values of K to get

$$P(F = f|\Gamma, \Lambda) = P(k|n, \Gamma, \Lambda) = \sum_{K=0}^{N} P(k|n, K, \Gamma, \Lambda)P(K|n, \Gamma, \Lambda). \tag{7.6}$$

Assuming that the subsample of n individuals was taken at random, then the number of observed carriers k given the total number of carriers K is independent of the particular inheritance path Γ, and follows the hypergeometric distribution:

$$P(k|n, K, \Gamma, \Lambda) = P(k|n, K) = H(k; N, n, K)$$

and similarly the true number of carriers is independent of the sampling:

$$P(K|n, \Gamma, \Lambda) = P(K|\Gamma, \Lambda)$$

giving

$$P(F = f|\Gamma, \Lambda) = P(k|n, \Gamma, \Lambda) = \sum_{K=0}^{N} H(k; N, n, K)P(K|\Gamma, \Lambda) \tag{7.7}$$

which we use in the calculation of (2.5) in the main text.

7.1.4 Regional Allele Frequency Estimates

We can use the simulated inheritance paths to estimate regional allele frequencies given the observed event S that the set of haplotypes c in the carrier individuals do indeed carry the minor allele, and the event that we observe f carriers unconnected to the genealogy, under the assumption Λ that Γ climbs from carriers of the minor allele and coalesces to a single individual within the genealogy. Letting R_m be the number of carriers in some subset of individuals m (usually defined as a geographic region), we have

$$E[R_m|F = f, \Lambda, S] = \sum_{r_m} r_m P(R_m = r_m|F = f, \Lambda, S). \tag{7.8}$$

Summing over all inheritance paths Γ, the chain rule gives

$$\begin{aligned} P(R_m = r_m|F = f, \Lambda, S) &= \sum_{\Gamma} P(R_m = r_m, \Gamma|F = f, \Lambda, S) \\ &= \frac{\sum_\Gamma P(\Lambda, R_m = r_m, \Gamma, F = f|S)}{P(F = f, \Lambda|S)} \\ &= \frac{\sum_\Gamma P(\Lambda|R_m = r_m, \Gamma, F = f, S)P(R_m = r_m|\Gamma, F = f, S)P(F = f|\Gamma, S)P(\Gamma)}{P(F = f, \Lambda|S)}. \end{aligned} \tag{7.9}$$

Where the last line uses the fact that $P(\Gamma|S) = P(\Gamma)$. Because the coalescence condition Λ is fully determined by Γ and S, we can write $P(\Lambda|\Gamma, S, \cdot) = P(\Lambda|\Gamma) = \mathbb{1}_\Lambda(\Gamma)$, where $\mathbb{1}_\Lambda(\Gamma)$ indicates whether Γ coalesces to a single lineage. Using the law of total probability and the chain rule on the denominator as well, we can write

$$P(R_m = r_m|F = f, \Lambda, S) = \frac{\sum_\Gamma \mathbb{1}_\Lambda(\Gamma)P(R_m = r_m|\Gamma, F = f, S)P(F = f|\Gamma, S)P(\Gamma)}{\sum_{\Gamma'} \left[\mathbb{1}_\Lambda(\Gamma')P(F = f|\Gamma', S)P(\Gamma')\right]}. \tag{7.10}$$

We can now write (7.8) as

$$E[R_m|F = f, \Lambda, S] = \sum_{r_m} r_m \frac{\sum_\Gamma \mathbb{1}_\Lambda(\Gamma)P(R_m = r_m|\Gamma, F = f, S)P(F = f|\Gamma, S)P(\Gamma)}{\sum_{\Gamma'} \left[\mathbb{1}_\Lambda(\Gamma')P(F = f|\Gamma', S)P(\Gamma')\right]} \tag{7.11}$$

$$= \frac{E_\Gamma\left[\mathbb{1}_\Lambda(\Gamma)E[R_m|\Gamma, F = f, S]P(F = f|\Gamma, S)\right]}{E_\Gamma\left[\mathbb{1}_\Lambda(\Gamma)P(F = f|\Gamma, S)\right]}. \tag{7.12}$$

We then estimate $P(F = f|\Gamma, S)$ using the methods described in the main text and Appendix 7.1.3.

Computing $E[R_m|\Gamma, F = f, S]$ is challenging, because we do not have an expression for the distribution of R_m conditioning on F. We do have an expression for $E[R_m|\Gamma, S]$, but R_m is not independent of f: when performing allele dropping from Γ, each transmission of the minor allele increases both the expectations of f and R_m.

To account for this correlation, we wish to simply scale the distribution based on the difference between the observed and expected global allele frequency. This is especially justified in a growing population, where an early success in allele transmission has a much larger effect on the variance of F and R_m than a later transmission. For example, if the founder individual transmits the minor allele to 8 out of 8 offspring, the expected descendant allele frequency among descendants is double its naive expectation. By contrast, the same information about a recent individual who is only one among hundreds of carriers will only have a marginal effect on the expected frequency. We can therefore consider that the global allele frequency is a random variable that is primarily determined by the proportion σ of individuals in $\partial\Gamma$ who receive the minor allele, and neglect the subsequent variation. If the sample size n is large enough, the allele frequency F drawn from a given inheritance path Γ is approximately $2\sigma e_\Gamma$, where e_Γ is the expected allele frequency generated from Γ.

Under this simplified model, we can compute

$$\begin{aligned}
E[R_m|\Gamma, F = f, S] =& \sum_{r_m} r_m P(R_m = r_m|\Gamma, F = f, S) \\
=& \sum_{r_m} r_m \sum_{\sigma} P(R_m = r_m, \sigma|\Gamma, F = f, S) \\
=& \sum_{r_m} r_m \sum_{\sigma} P(R_m = r_m|\sigma, \Gamma, F = f, S) P(\sigma|\Gamma, F = f, S) \\
=& \sum_{r_m} r_m \sum_{\sigma} P(R_m = r_m|\sigma, \Gamma, F = f, S) \delta_{\sigma - \frac{F}{2e_\Gamma}} \\
=& \sum_{r_m} r_m P(R_m = r_m|\sigma = \frac{F}{2e_\Gamma}, \Gamma, F = f, S) \\
=& \sum_{r_m} r_m P(R_m = r_m|\sigma = \frac{f}{2e_\Gamma}, \Gamma, S) \\
=& E[R_m|\sigma = \frac{f}{2e_\Gamma}, \Gamma, S].
\end{aligned} \tag{7.13}$$

Since $R_m \simeq \sum_{i\in\partial\Gamma} B_{m,i}$, where $B_{m,i}$ is the number of minor alleles inherited, in population m, from boundary individual i, we find $E[R_m|\sigma, \Gamma, S] \simeq \sum_{i\in\partial\Gamma} E[B_{m,i}] = \sum_{i\in\partial\Gamma} \sigma E[C_{m,i}]$, where $C_{m,i}$ is the number of minor alleles inherited, in population m, from boundary individual i, conditional on i carrying a minor allele. Since $E[R_m|\Gamma, S] = \sum_{i\in\partial\Gamma} \frac{1}{2} E[C_{m,i}]$, we conclude $E[r_m|\sigma] \simeq 2\sigma E[r_m]$, and

$$E[R_m|\Gamma, F = f, S] = \frac{f}{e_\Gamma} E[R_m|\Gamma, S]. \tag{7.14}$$

In other words, we rescale the expected allele regional frequencies by the ratio of predicted to observed global allele frequencies.

Using the importance sampling scheme described in the main text to simulate only those Γ_j which coalesce to a single founder, implying that $\mathbb{1}_\Lambda(\Gamma_j) = 1$ for all $i = 1, ..., M$, the expected regional allele frequency estimate becomes:

$$E[R_m|F = f, \Lambda, S] \simeq \frac{f}{e_\Gamma} \frac{\sum_{j=1}^{M} \frac{p(\Gamma_j)}{q(\Gamma_j)} E[R_m|\Gamma, S] P(F = f|\Gamma, S)}{\sum_{j=1}^{M} \frac{p(\Gamma_j)}{q(\Gamma_j)} P(F = f|\Gamma, S)}. \tag{7.15}$$

Kinship-Based Regional Allele Frequency Estimates

Since calculating all pairwise kinship scores for probands of the BALSAC genealogy would require generating a matrix with the order of 10^{12} entries, we take a random sample of 100 probands from each of 23 geographic regions of Quebec. Then for each simulated patient panel, we calculate the average kinship of these groups of 100 individuals with all patients.

Note that the approximation in (7.14) guarantees that our estimate of the global allele frequency is always exactly equal to the observed allele frequency. To ensure a fair comparison when evaluating the accuracy of importance sampling versus kinship-based methods, we use a similar scaling factor to incorporate the global allele frequency information into kinship estimates. Denoting regional mean kinship estimates by $\hat{r}_{m,kin}$ and the global mean kinship estimate by $\hat{f}_{kin}$, we use the estimator

$$\hat{r}_{m,\text{kin, corrected}} = \hat{r}_{m,\text{kin}} \frac{f}{\hat{f}_{\text{kin}}}$$

to calculate our kinship-based regional estimates.

7.1.5 Validating the Calibration of Ancestor Posterior Probabilities

As described in the main text, we validate the posterior probabilities of groups of ancestors within relatedness clusters. Relatedness clusters are defined as groups of ancestors who together have only a single shared path to all carriers of the affected alleles. Each nuclear family group within such a cluster may have a single extra path to some carriers, as long as they have only a single path to all of them. Probabilities for cluster J are then given by:

$$P(A \in J|S) = \sum_{a_i \in J} P(A = a_i|S).$$

After generating validation panels and calculating the posterior probabilities for each relatedness cluster, we bin clusters by their posterior probability, and model the number of true generating ancestors in bin i as a binomial process with success probability ϕ_k. To generate confidence interval on ϕ_k, we let τ_k represent the total number of ancestors bin i and x_k the number of true generating ancestors. Assuming a flat prior for all ϕ_k,

$$P(\hat{\phi}_k|\tau_k, x_k) \sim \text{Beta}(x_k + 1, \tau_k - x_k + 1). \tag{7.16}$$

7.1.6 CAID Data and IBD Computation

11 homozygous patients were previously diagnosed and genetically characterized using the Illumina HumanOmni5-Quad chip [71]. We also used genotypes [72, 73, 74] from the Quebec Regional Population Sample (QRS) as a control group [75]. Among the 229 genealogically connected controls we found one heterozygous carrier of the CAID mutation, based on genotype and confirmed by Sanger sequencing. The observation of 3 carriers in a cohort of 900 genotyped French Canadians, from CARTaGENE [76], gave us our estimate of the CAID allele frequency.

Our assumption of a single origin for the CAID allele within the BALSAC genealogy is based on the sharing of a 2.9Mb homozygous segment on chromosome 3, described in the Applications section of the main text. This segment was discovered by analyzing segments within the patients which were Identical-By-Descent (IBD). The 11 patients and 229 control individuals gave 240 genotypes with which to evaluate the extent of pairwise IBD sharing. IBD was inferred by the analysis of more than 300 000 genotyped SNPs common to the patients and QRS controls, using BEAGLE 4 software [77].

7.2 Chapter 2 Supplemental Material

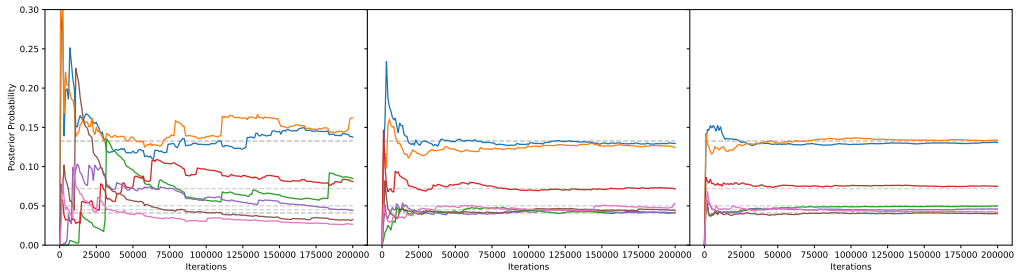


Figure 7.1: Convergence of likelihood estimates for 7 most-likely ancestors of a minor allele in a single simulated carrier panel. With importance sampling based left-to-right on: a possible path to coalescence only; the number of common ancestors shared with all other simulated carriers of the minor allele; likelihood of coalescing with other lineages.

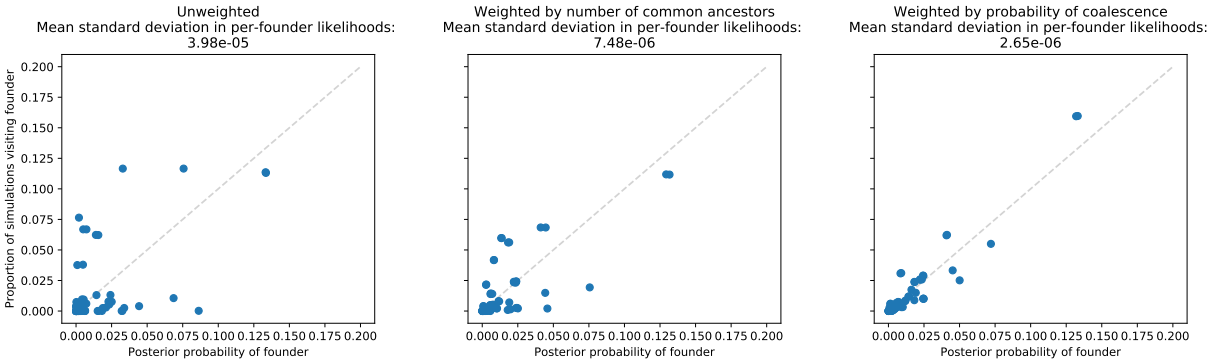


Figure 7.2: Proportion of simulated inheritance paths which lead to each founder versus converged founder posterior probability. With importance sampling based left-to-right on: a possible path to coalescence only; the number of common ancestors shared with all other simulated carriers of the minor allele; likelihood of coalescing with other lineages. Uses the same simulated carrier panel as Fig. 7.1. Importance sampling convergence is fastest when outcomes are sampled proportionally to their true probability [78].

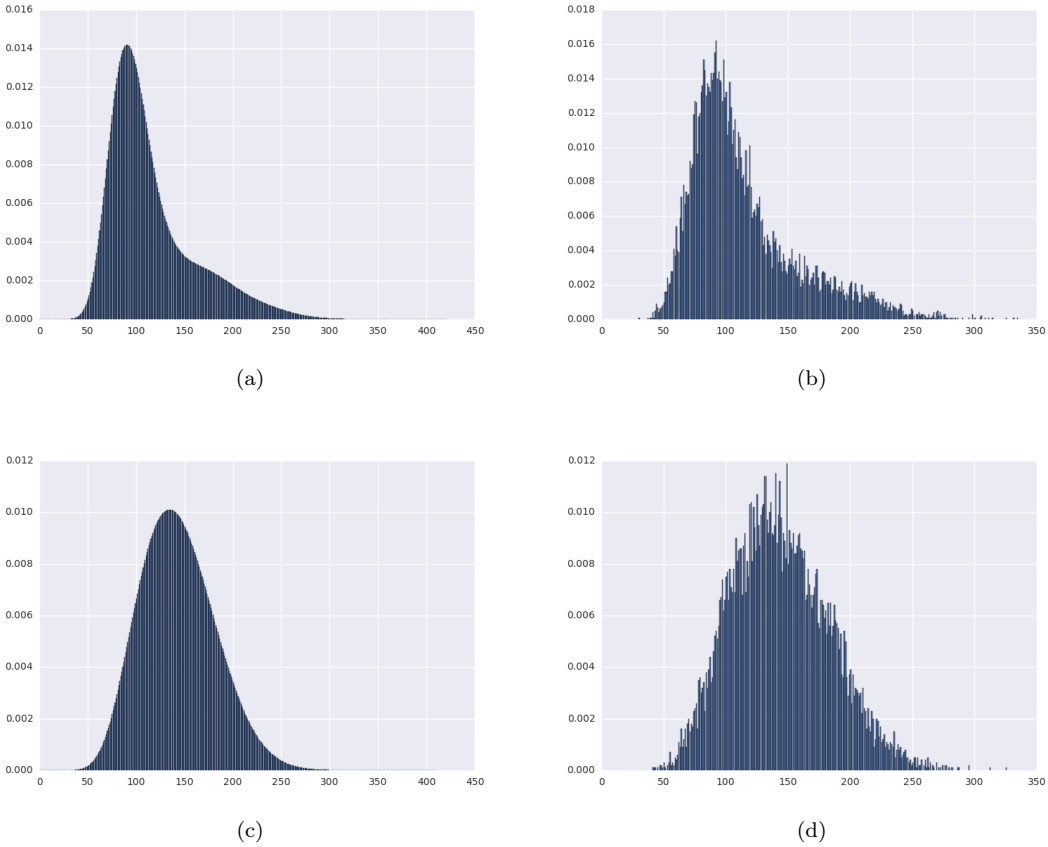


Figure 7.3: Comparison of simulated inheritance path allele frequency distributions (B, D) and their approximation via convolution of the distributions of the tree boundary (A, C) using the method described in Appendix 7.1.3.

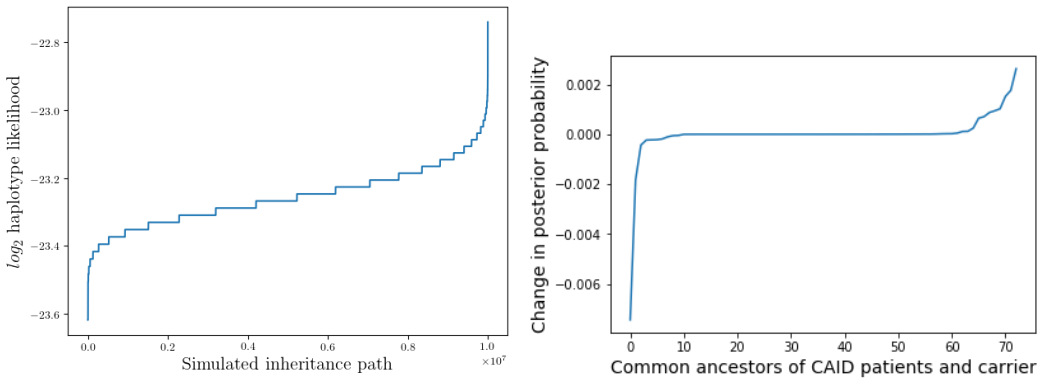


Figure 7.4: (A) Log-likelihoods of observing shared 2.9Mb segment in CAID patients and carrier, over all simulated inheritance paths. (B) Impact of incorporating shared haplotype length among CAID patients on estimated posterior probabilities of each common ancestor having been the true origin of the minor allele.

Ind	Father	Mother	Sex
1	11	12	1
2	15	14	2
3	15	14	2
11	102	101	1
12	0	0	2
13	102	101	1
14	0	0	2
15	103	104	1
16	103	104	2
18	105	106	2
19	105	106	2
20	107	108	2
21	107	108	1
101	0	0	2
102	202	201	1
103	0	0	1
104	202	201	2
105	202	201	1
106	0	0	2
107	202	201	1
108	0	0	2
201	0	0	2
202	0	0	1

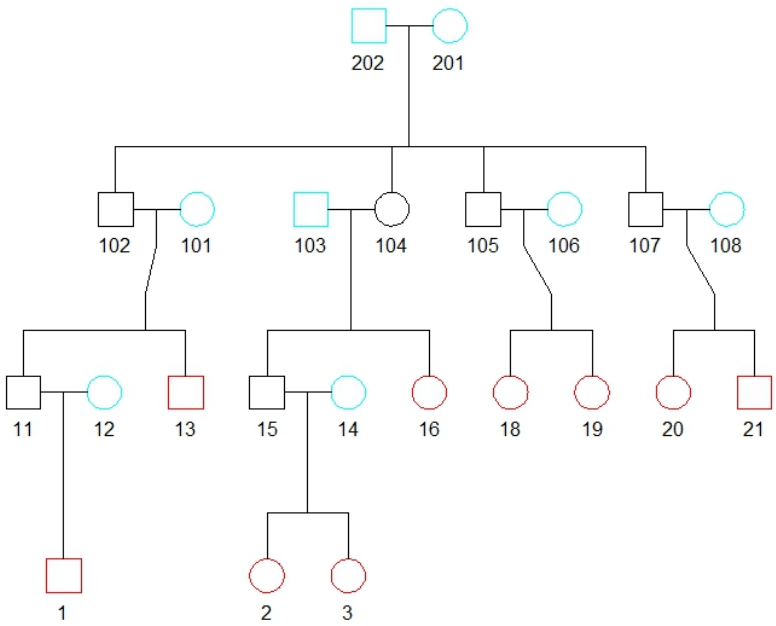


Table 7.1: Example pedigree and corresponding data format.

Region	Years	Baptisms	Marriages	Deaths	Total
Abitibi	1898-1985	15	19 210		19 225
Bas-Saint-Laurent	1701-1985	14	85 606		85 620
Beauce	1740-2013	17	58 515	3	58 535
Bois-francs	1671-2008	38	128 656		128 694
Charlevoix	1686-1995	91 380	29 614	48 410	169 404
Côte-de-Beaupré	1661-1984	1	19 803		19 804
Côte-du-Sud	1679-1985	6	87 223		87 229
Côte-Nord	1677-2002	6	16 549		16 555
Estrie	1781-1989	14	128 648		128 662
Gaspésie	1693-1984	5	45 221		45 226
Ile de Montréal	1643-2001	69	529 652		529 721
Iles de la Madeleine	1772-1991	9 108	5 942	2 410	17 460
Lanaudière	1672-2007	8	93 409		93 417
Laurentides	1690-2003	4	61 060		61 064
Mauricie	1645-2002	115	111 673		111 788
Outaouais	1806-1993	66	111 872		111 938
Québec (agglomération)	1621-2007	18	155 656		155 674
Région de Québec	1675-2006	8	83 294		83 302
Reste du Québec	1936-1985		1 708		1 708
Richelieu	1668-2006	9	148 571		148 580
Rive nord ouest (Montréal)	1679-1992	3	63 297		63 300
Rive sud (Montréal)	1670-1985	5	70 880		70 885
Saguenay-Lac-St-Jean	1833-2007	431 464	92 721	122 959	647 144
Témiscaminque	1881-1984	14	14 199		14 213
Lieu indéterminé (au Québec)	1657-2006		1 425		1 425
ENSEMBLE DU QUÉBEC		431 495	2 164 404	173 782	2 870 573

Source : Fichier BALSAC, August 31 2017

* All records prior to 1800 (N=69 000) come from the Programme de recherche en démographie historique (PRDH) of University of Montreal. They were obtained through exchanges and collaborative arrangements.

Figure 7.5: Number of vital event records per region of Quebec [37]. Table reproduced July 18th, 2018 from `http://balsac.uqac.ca/english/balsac-database/overview-of-data/`

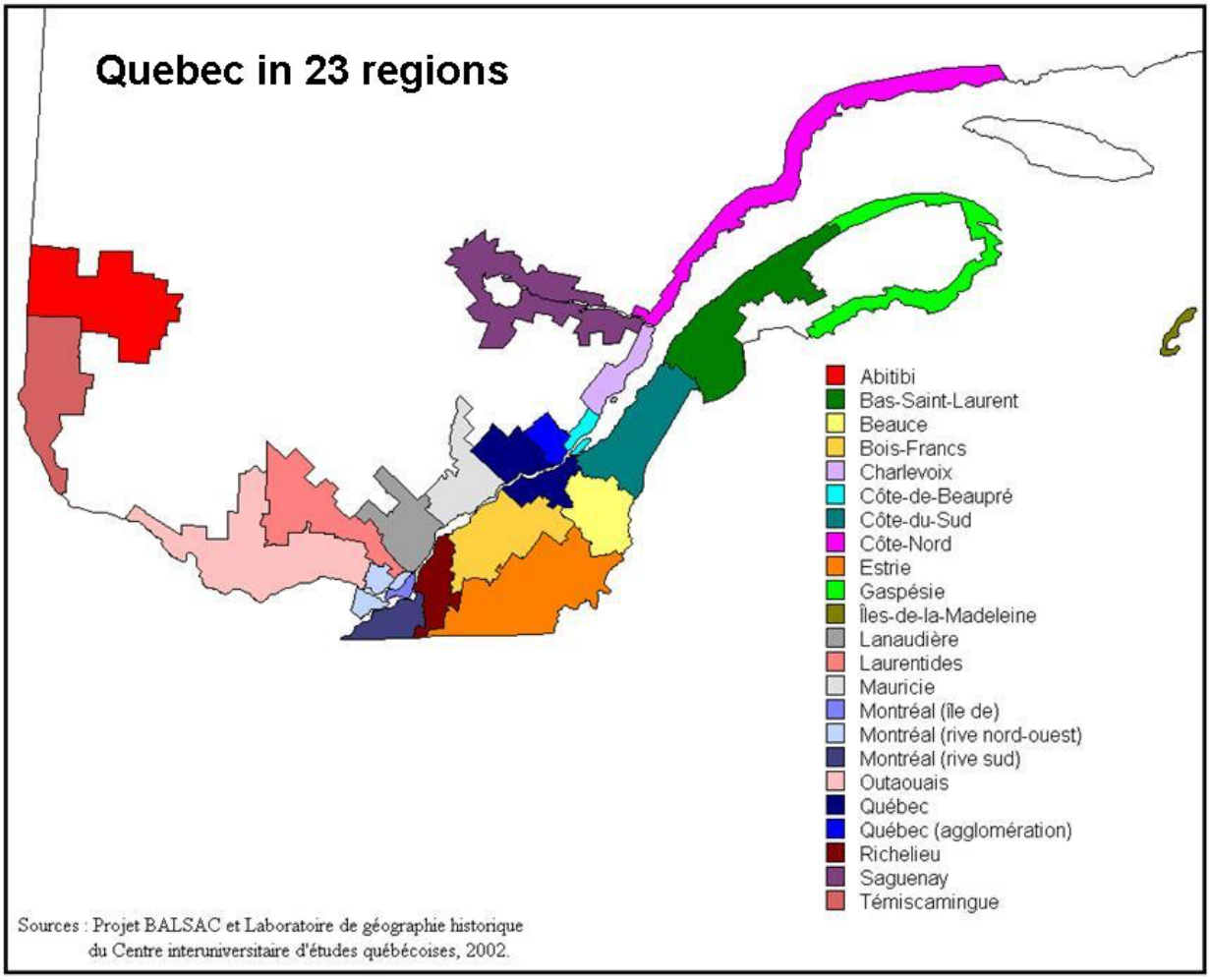


Figure 7.6: Quebec regions used in the BALSAC project [37]. Figure reproduced September 12th, 2018 from `http://balsac.uqac.ca/english/balsac-database/overview-of-data/`

Region	Estimated Allele Frequency	95% Confidence Interval
ABITIBI	0.00128	(0.00127, 0.00129)
BAS SAINT LAURENT	0.00163	(0.00156, 0.00169)
BEAUCE	0.00425	(0.00408, 0.00443)
BOIS FRANCS	0.000882	(0.000858, 0.000908)
CHARLEVOIX	0.00643	(0.00630, 0.00654)
COTE DE BEAUPRE	0.00417	(0.00410, 0.00423)
COTE DU SUD	0.00183	(0.00176, 0.00190)
COTE NORD	0.00253	(0.00249, 0.00258)
ESTRIE	0.00144	(0.00141, 0.00146)
GASPESIE	0.000738	(0.000696, 0.000767)
ILE DE MONTREAL	0.000588	(0.000580, 0.000596)
ILES DE LA MADELEINE	2.61e-05	(2.45e-05, 2.80e-05)
LANAUDIERE	0.000462	(0.000450, 0.000473)
LAURENTIDES	0.000500	(0.000486, 0.000515)
MAURICIE	0.000808	(0.000789, 0.000825)
OUTAOUAIS	0.000349	(0.000340, 0.000356)
QUEBEC (AGGLOMERATION)	0.00183	(0.00179, 0.00187)
REGION DE QUEBEC	0.00118	(0.00113, 0.00124)
RICHELIEU	0.000581	(0.000566, 0.000598)
RIVE NORD OUEST (MTL)	0.000390	(0.000382, 0.000399)
RIVE SUD (MTL)	0.000477	(0.000470, 0.000482)
SAGUENAY (LAC ST JEAN)	0.00520	(0.00512, 0.00527)
TEMISCAMINGUE	0.000794	(0.000786, 0.000802)
All Probands	0.00167	

Table 7.2: Estimated regional frequencies of the CAID allele within the province of Quebec, among individuals linked to the BALSAC genealogical database. Confidence intervals estimated from bootstrapping over simulated inheritance paths.

Error Measure	Kinship-Based	ISGen	ISGen / Kinship
MAE	0.00105	0.000784	0.74
RMSE	0.00204	0.00169	0.83
MAE (estimated freq < 0.005)	0.000885	0.000591	0.67
RMSE, (estimated freq < 0.005)	0.00146	0.000983	0.67

Table 7.3: Mean absolute error and root mean squared error in regional allele frequency estimates for ISGen (path-based) and a kinship-based method. We simulated 100 patient panels and corresponding regional allele frequencies. Simulated regional allele frequencies were compared to inference results based on patient panels and estimated global allele frequency.

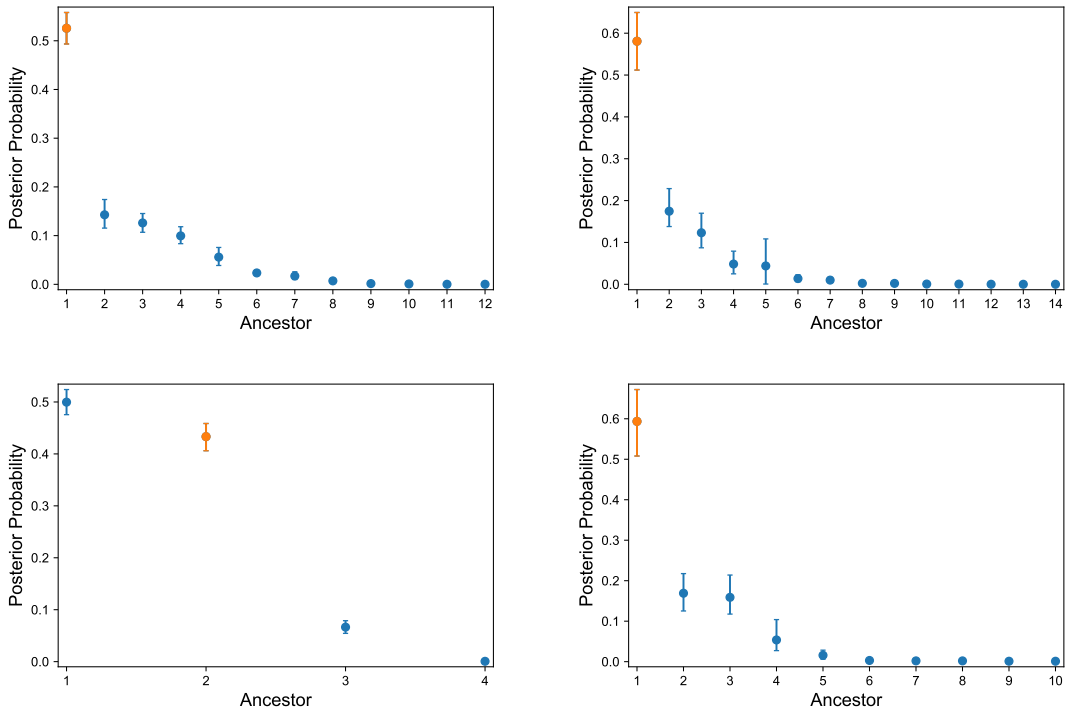


Figure 7.7: Ancestor posterior probabilities for 4 simulated patient panels, similar to the one displayed in Figure 2.4. The ancestor generating the panel is shown in orange. Error bars represent uncertainty due to the finite sample size (i.e., the finite number of iterations) in importance sampling. 95% confidence intervals were obtained from bootstrapping over iterations. Only ancestors with nonzero posterior probability are displayed, and ancestor labels represent ordering by posterior probability for a given simulation.

7.3 Chapter 3 Appendices

7.3.1 S1 Appendix. Wright-Fisher Implementation Details

We describe here the precise order of events happening at each generation in our implementation of the Wright-Fisher model. For more technical details see the documentation at `https://msprime.readthedocs.io`, as well as the source code at `https://github.com/tskit-dev/msprime`.

In the first ('current') generation, samples are initialized as haploid copies of the region to be simulated (which can later be paired to form diploid individuals). Lineages of each sample are then constructed backwards in time as follows (detailed comments labelled by pseudocode line number are provided below):

Algorithm 1 Wright-Fisher simulations in `msprime`

```
1: time ← 0
2: while number of extant lineages > 0 do
3:     time ← time + 1
4:     migrate lineages (migration rates, time)
5:     apply demographic events (time)
6:     choose parents for all extant lineages
7:     recombine extant lineages
8:     record coalescence events
```

4 Migration events are drawn according to the forwards-time rates provided, and migrant lineages are moved to their new population. This is equivalent to migration of gametes, as opposed to migration of diploid individuals. A forwards-time event from population i to j moves a lineage from population j to i backwards in time.

5 Demographic events are carried out, such as changes to population sizes or growth rates, mass migrations, or bottlenecks.

6 Each haploid lineage draws a diploid parent within its current population.

7 Recombination occurs, with each breakpoint alternately assigning segments to be inherited from one of the two parental copies of the genome (back-and-forth recombination, see Fig 3.1 in the main text).

8 Segments inheriting from the same parental copy of the genome are merged into a single lineage, with coalescent events recorded in overlapping regions.

When there is a single ancestral lineage at every position in the simulated genome, the simulation terminates.

Our whole-genome simulations are performed with a single chromosome of length 35.13 Morgans and 22 'effective' chromosomes of realistic lengths separated by 0.5 Morgans. This is not exactly equivalent

to simulating fully independent chromosomes. However, this should not have a qualitative impact on the analyses considered here.

7.3.2 S2 Appendix. Long-range linkage disequilibrium

For pairs of loci at low recombination distances ($r \ll 1$), it is unlikely for more than a single recombination event to occur in a given meiosis. In this case, the coalescent accurately models LD between linked loci. For larger recombination distances, loci only become unlinked under an odd number of recombinations. This has probability $P(\text{odd \# rec. events}|r) = \frac{1-e^{-2r}}{2}$, which has a maximum value of 1/2. This leads to non-zero long-range LD, even in the case of fully unlinked loci [79]. The diploid Wright-Fisher captures this, but coalescent estimates of LD decay to zero for increasing r (Fig 7.8).

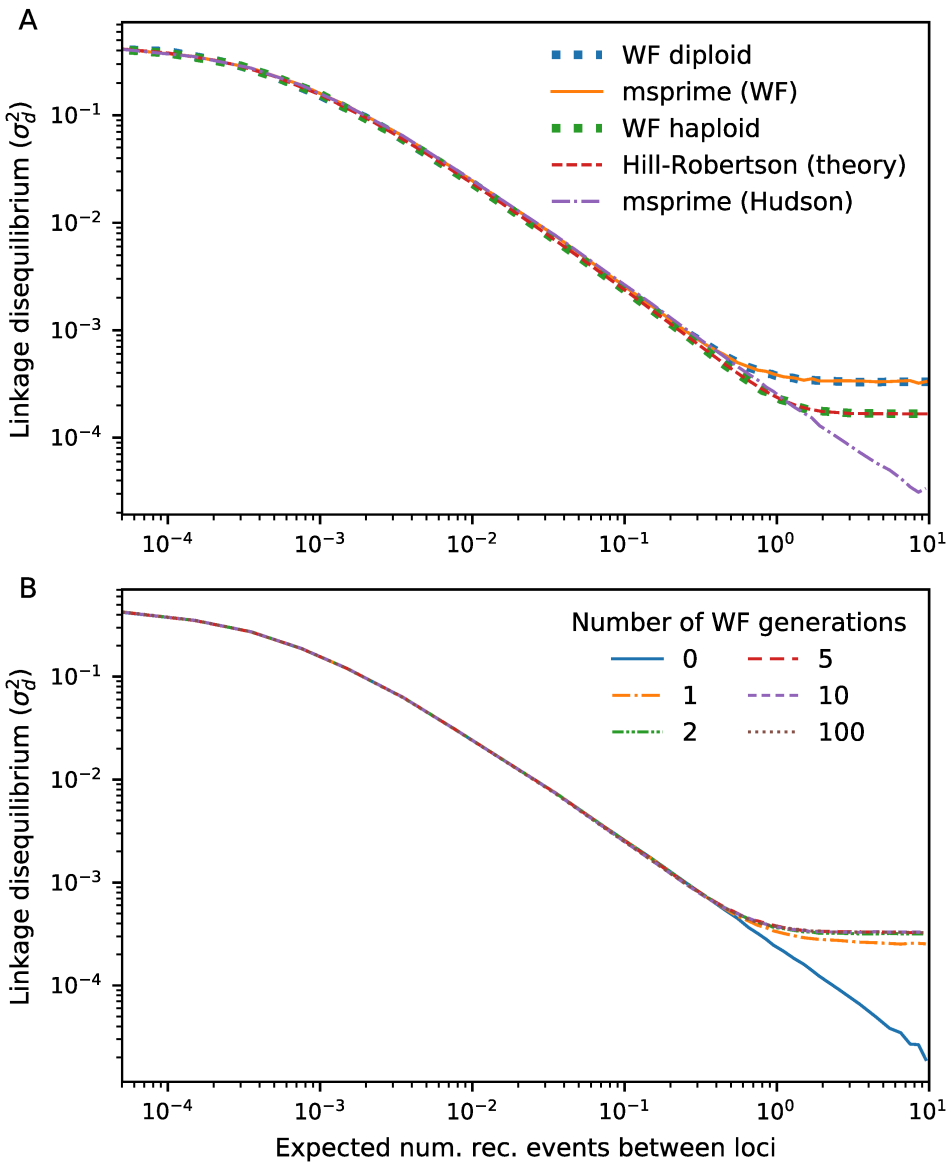


Figure 7.8: Linkage disequilibrium as measured by $\sigma_D^2 = E[D^2]/E[p(1-p)q(1-q)]$ under different simulation and theory models [80]. Simulations were carried out with population size $N = 1000$ at steady state demography for a single 10M chromosome. At fully unlinked loci, the expected value of σ_D^2 is $\frac{1}{3N}$ in a diploid model and $\frac{1}{6N}$ in a haploid model [81]. (A) Hudson and Wright-Fisher simulations. (B) Hybrid simulations with varying numbers of Wright-Fisher generations before switching to the Hudson coalescent.

7.3.3 S3 Appendix. An approximate model for IBD sharing

To provide a simple analytical model for the relationship between total length and counts of IBD sharing, we simply consider a pair of haploid samples sharing a single diploid common ancestor at time t generations in

the past and estimate the expected number and length of haplotypes shared given t (similar derivations have been made in [82, 83, 84, 85]). We can think of the ancestry of each haploid genome as a mosaic formed by copying genomic segments from its 2^{t-1} possible ancestors. Similarly, a pair of haploid samples can be seen as a mosaic formed by copying from one ancestor for each sample. We can define paired-ancestry segments as continuous segments having no changes in ancestry in either sample. By this definition, if each sample has K chromosomes of total length L Morgans, the pair will have on average $K + 2Lt$ paired-ancestry segments.

Since each haploid sample has 2^{t-1} possible ancestors from which to inherit genetic material, a pair of samples will both inherit a paired-ancestry segment from their common ancestor with probability $\frac{1}{2^{2t-2}}$. Since the ancestor is diploid, they inherit from the same ancestral copy of the genome with probability $\frac{1}{2}$. The probability that a paired-ancestry segment is IBD in the pair is therefore $\frac{1}{2^{2t-1}}$, and the expected number of IBD segments s between the pair is:

$$s = \frac{K + 2Lt}{2^{2t-1}}. \tag{7.17}$$

The length of the genome shared, denoted by x, corresponds to L times the probability of having a shared ancestor at any particular locus, which is $\frac{1}{2^{2t-1}}$, giving:

$$x = \frac{L}{2^{2t-1}}. \tag{7.18}$$

The expected values (s, x) are shown in Fig 3.3 in main text as white dots for t from 1 to 5 generations, corresponding to half-siblings, first half-cousins, and so on.

Under monogamy, a similar argument can be used to obtain

$$s = \frac{K + 2Lt}{2^{2t-2}}. \tag{7.19}$$

and

$$x = \frac{L}{2^{2t-2}}. \tag{7.20}$$

Similarly, avuncular relationships in monogamy have three meioses (so $K + 3L$ segments) with a sharing probability of $\frac{1}{2}$ with one of the two shared ancestors.

$$s = \frac{K + 3L}{2} \text{ and } x = \frac{L}{2}, \tag{7.21}$$

and grandparent-offspring have two meioses ($K + 2L$ segments) wth a sharing probability of $\frac{1}{2}$ between the

two haplotypes of the grandparent.

$$s = \frac{K + 2L}{2} \text{ and } x = \frac{L}{2}. \tag{7.22}$$

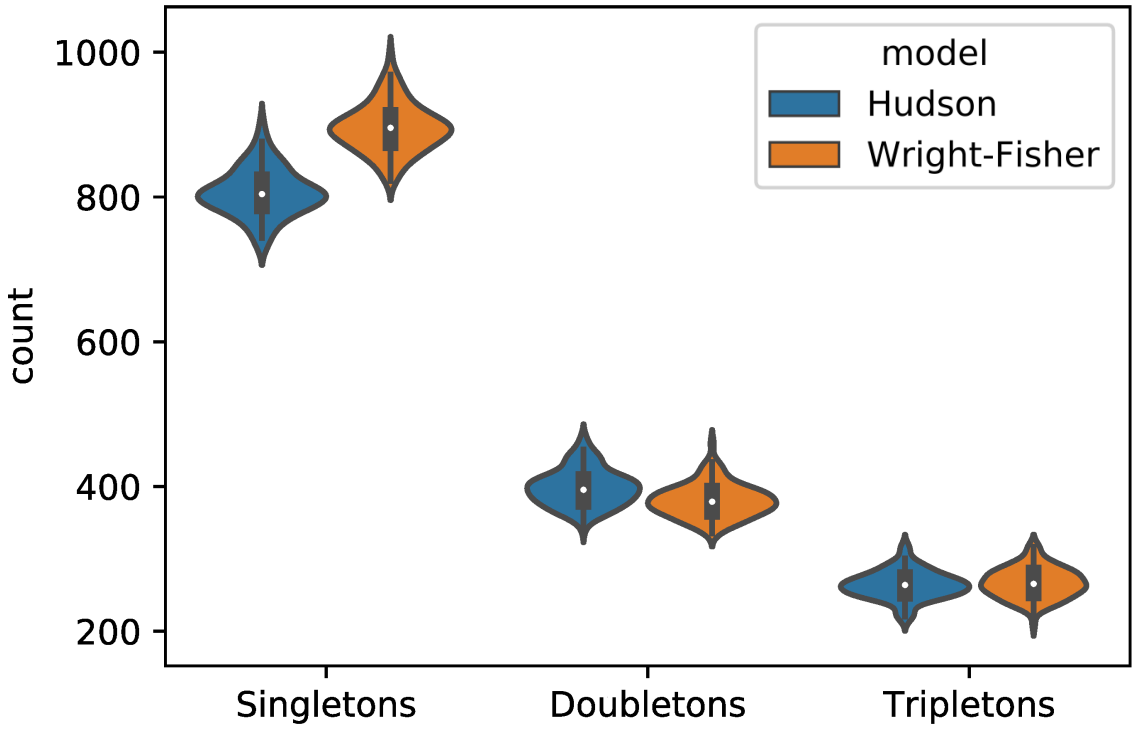
However, we expect that meioses occurring in the grandparent could be hidden by the statistical phasing, so that this result would be particularly sensitive to inaccurately phased data.

7.3.4 S4 Appendix. The Genizon Biobank

The Genizon Biobank is composed of 26 cohorts for complex diseases and one general control cohort, totalling 44,981 participants. It is administered by Genome Quebec and was originally collected by Genizon BioSciences Inc. Participants were all residents of the province of Quebec [47].

We calculated IBD using genotypes from all cohorts (except Asthma and Crohn disease which were genotyped with different chips), totalling 9,961 individuals (cases and controls) including trios and duos. We removed trios and duos keeping the parents when possible, as well as individuals with less than 99% genotype calls over all SNPs. This left us with 8,435 individuals genotyped for 233,927 SNPs (keeping SNPs with at least 99% genotypes over all individuals). This filtering was done on data from each cohort individually using plink version 1.90, and they were then merged. Phasing was done using shapeit version 2.r790 and IBD was estimated using GERMLINE 1.5.1, using a minimum segment length of 5 centimorgans. Very high IBD peaks on chromosomes 1 and 9 were removed using an in-house script.

7.4 Chapter 3 Supplemental Material

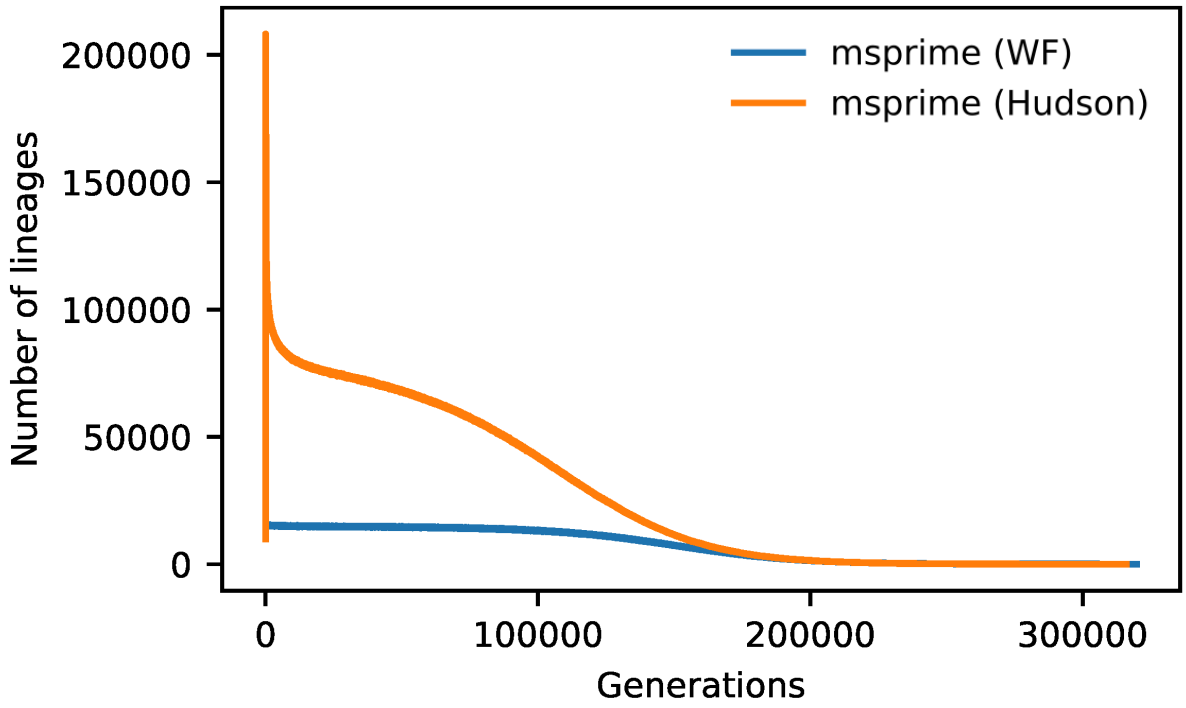


S1 Figure. Number of singletons, doubletons, and tripletons simulated under Wright-Fisher and Hudson coalescent models. A 1Mb region was simulated 100 times in 20,000 haploid lineages in a diploid population of 10,000 individuals.

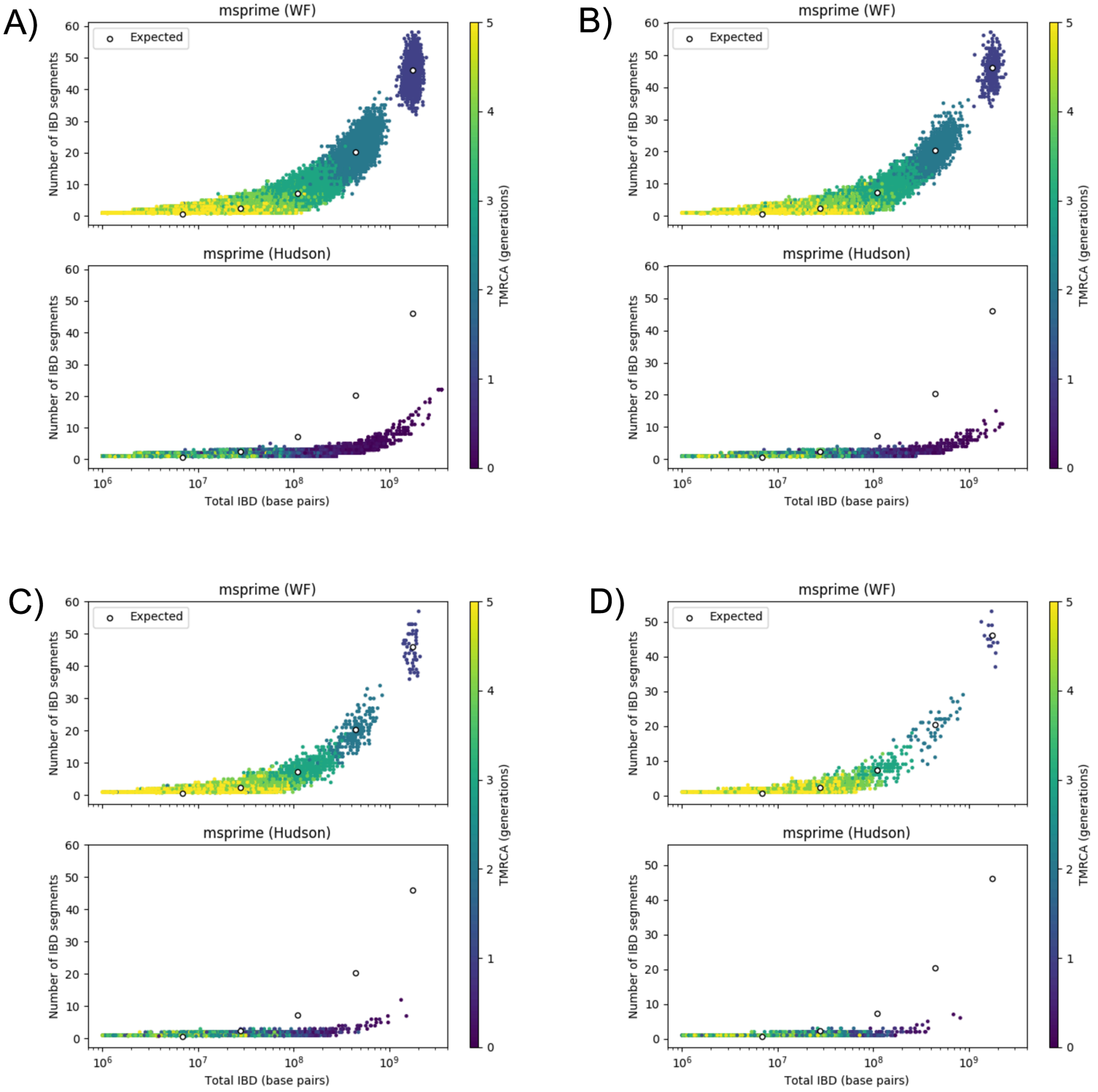
S1 Table. Relative difference in mean number of singletons, doubletons, and tripletons under the Wright-Fisher (N_{WF}) and Hudson (N_H) models.

Frequency	$\frac{N_{WF}-N_H}{N_{WF}}$
Singletons	0.099131
Doubletons	-0.047253
Tripletons	0.010092

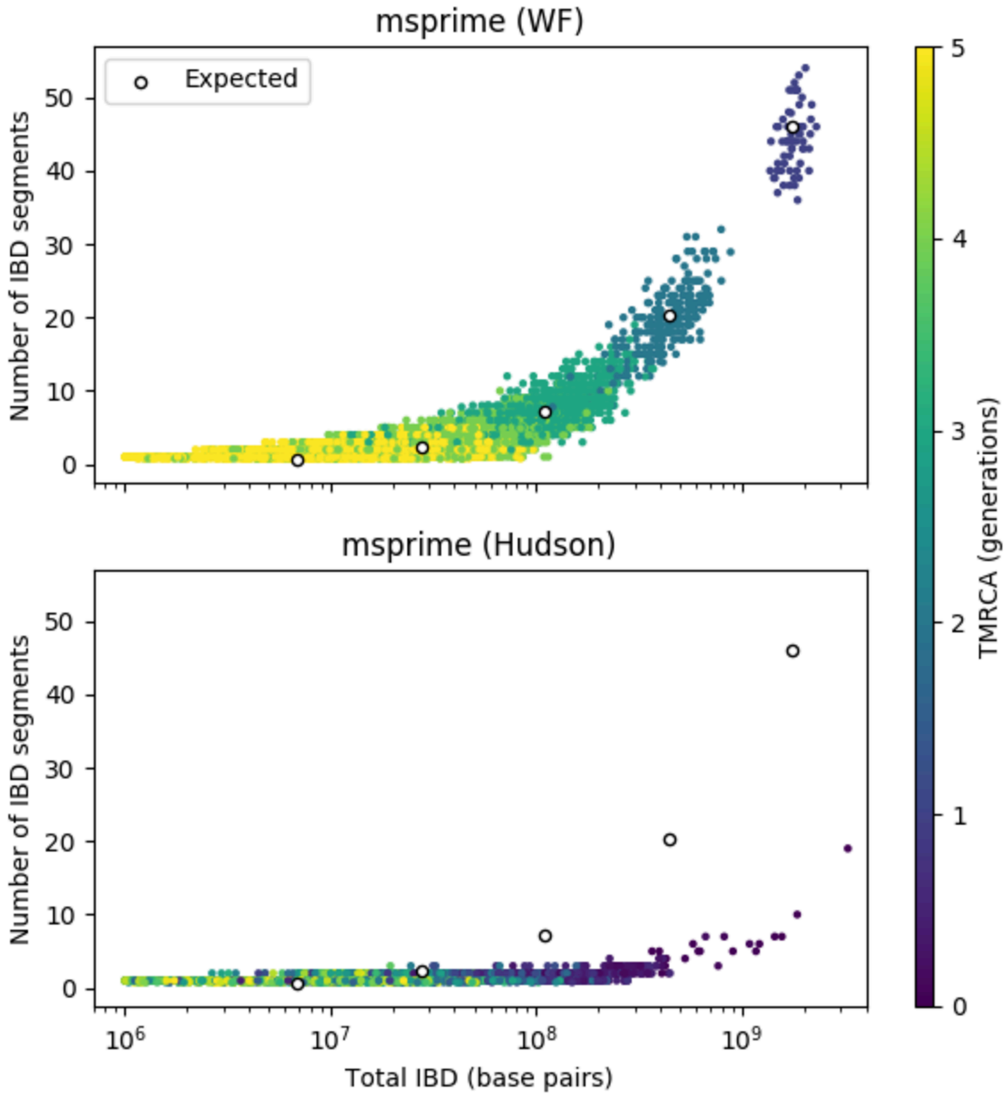
From data shown in S1 Figure. These results closely match those presented in [67, 86].



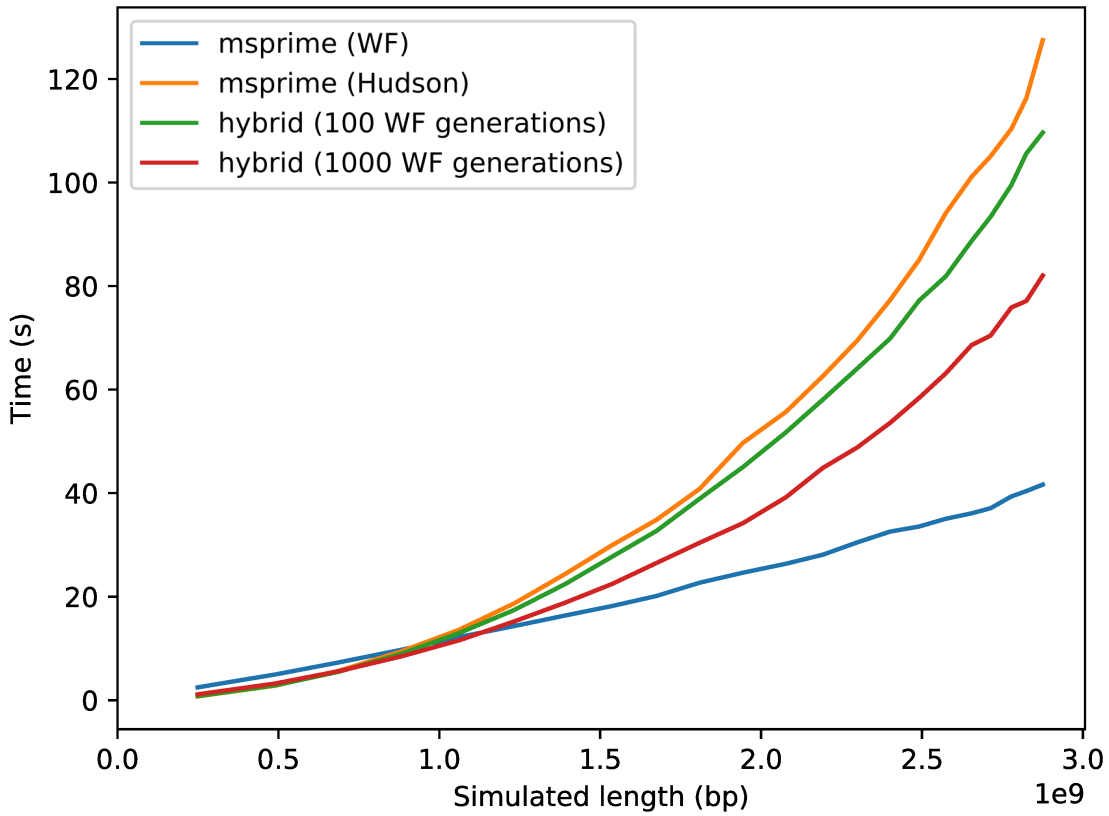
S2 Figure. Number of surviving lineages over time in coalescent and back-in-time Wright-Fisher dynamics. We simulated 10,000 haploid whole genomes with 22 chromosomes of realistic lengths in a population of 10,000 diploid individuals. The method for simulating multiple chromosomes is described in S1 Appendix. Similar results were shown in [68].



S3 Figure. Number of IBD segments between pairs of individuals versus total length of shared IBD segments. 22 chromosomes of realistic lengths, simulated under Wright-Fisher model (top) and coalescent (bottom), compared to the analytical expectation under Eqs (1) and (2) of S3 Appendix. Effective population size 10,000, sample size A) 5000, B) 2500, C) 1000, D) 500. Minimum IBD segment length of 1 centimorgan.



S4 Figure. Number of IBD segments between pairs of individuals versus total length of shared IBD segments, under the Gutenkunst et. al. (2009) [69] out-of-Africa model. 22 chromosomes of realistic lengths, simulated under Wright-Fisher model (top) and coalescent (bottom), compared to the analytical expectation under Eqs (1) and (2) of S3 Appendix. The African, European, and Asian populations had 1000 haploid samples each.



S5 Figure. Computation time of Hudson coalescent, Wright-Fisher, and hybrid models with 100 and 1000 Wright-Fisher generations before switching to the coalescent. Simulations contain from 1 to 22 chromosomes of realistic lengths, using the method described in S1 Appendix, in 500 haploid samples within a diploid population of size 500.

7.5 Chapter 4 Appendices

7.5.1 Pedigree Simulation Algorithm

Algorithm 2 Pedigree simulations in `msprime`

assign genomic segments to each individual
ind_heap ← *heap queue of individuals to simulate*
while *number of extant lineages > 0* and ind_heap *is not empty* **do**
    next_individual ← *pop most recent individual from* ind_heap
    time ← *time of* next_individual
    merge ancestral segments in next_individual, *recording coalescence events*
    recombine ancestral segments in next_individual
    if *parents of* next_individual *in pedigree* **then**
        assign recombined segments to parents of next_individual
        add parents of next_individual *to* ind_heap
    else
        add recombined segments to randomly-mating population
continue normal Wright-Fisher simulations of remaining ancestral segments